Information Allocation and Prosodic Expressiveness in Continuous Speech: A Mandarin Cross-Genre Analysis

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

[1] Phonetics Lab, Institute of Linguistics, Academia Sinica Taipei, Taiwan
[2] Taiwan International Graduate Program (TIGP), Academia Sinica Taipei, Taiwan
cytling@sinica.edu.tw

ABSTRACT

In addition to discourse association and assuming that allocation of key information is an important feature of prosodic expressiveness of continuous speech, the common accentuation patterns across 3 Mandarin speech genres through 4 degrees of perceived emphases are derived. Using frequency count as another control, it is found that only 6 types of emphasis patterns are needed account for 70% of the speech data regardless of genre. The 6 emphasis types are further compared for the distribution of (1) discourse units and emphasis tokens by speech genre, (2) emphasis pattern by phrase and (3) with respect to discourse positions to see if genre-specific features could be found. Results reveal that genre-dependent features can also be accounted for. In addition, individual genre properties are found to also be correlated with phrase length and specific emphasis patterns.

Index Terms—information allocation, emphasis pattern, prosodic expressiveness, speech genre

1. INTRODUCTION

Our previous studies on Mandarin spoken discourse prosody have demonstrated that cross-phrase discourse association is an important component of the prosody of continuous speech. Multi-phrase discourse unit featuring pitch reset of the initial phrase, flattened contours of the medial phrases and declination to terminal fall of the final phrase. Contributions from the cross-phrase associative pattern trigger systematic and accountable compensations of the lower level discourse units. Hence syllable tones, word contours and individual phrase intonation are modified in accordance with their respective discourse positions, and cumulative contributions account for approximately 75% of F0 output of multi-phrase discourse unit [1]. Overall fast-to-slow phrase tempo also contributes and account for overall tempo output in similar manner [2, 3]. But of course there are additional prosodic contributions other than phrase association that makes continuous speech more expressive. While the remaining make-up of the prosody of continuous speech may largely be attributed to qualitative contributions from speaker attitude and emotion, we believe another important linguistic factor that contributes to output prosody expressiveness is to signal key information in the speech flow through accentuation or emphasis. By accentuation/emphasis we refer to perceived prominence of words that stand out from their context [4, 5] to express stress and or focal points. Assuming that allocation of key information is an important feature of prosodic expressiveness, we will try to derive the most frequently used common accentuation patterns across speech genre as a control to further sort out genre-specific components. At the same time, we will also attempt to account for individual genre properties in relation to discourse structure [6]. In the following study, we will compare the distribution of the most frequently used emphasis patterns across three genres of Mandarin continuous speech in order to see how information allocation may affect overall output prosody and how genre related features can be found.

2. SPEECH MATERIALS AND ANNOTATION RATIONALE

2.1 Speech Materials

The materials used are microphone speech of three genres: (1) passive reading of 26 discourse pieces produced by 1 female radio announcer (45 min/11594 syllables/85MB, coded CNA) [7], (2) semi-active reading simulating weather broadcast produced by 1 female untrained speaker (approximately 45 min/7061 syllables/50MB, coded WB), and (3) spontaneous classroom lecture produced by 1 male university professor (approximately 26 min/7660 syllables/49 MB, coded LEC).

2.2 Annotation and rationale

Preprocessing is force-aligned segments by the HTK Toolkit followed by manual spot-checking. The speech data were manually tagged for 5 levels of perceived discourse boundaries [1, 2 and 2.1.1], as well as 4 levels of perceived emphases (see 2.2.2). Independent tagging by discourse units and perceived emphasis enables examination of interaction between perceived emphasis and paragraph/discourse structure.

2.2.1. Tagging perceived discourse boundaries and discourse positions

The speech data are manually tagged by trained transcribers into the HPG discourse units [1, 2]. The hierarchical HPG framework specifies 5 levels of perceived discourse prosodic boundaries B1 through B5. Corresponding discourse prosodic units are defined by chunks located inside each level of boundary breaks, namely, the syllable (SYL), prosodic word (PW), prosodic phrase (PPh), breath group (BG, a physio-linguistic unit constrained by change of breath while speaking continuously) and multiple -phrase speech paragraph PG. By default the relationship between the boundary breaks and prosodic units can be expressed as SYL/B1<PW/B2<PPh/B3<BG/B4<PG/B5 [8, 9]. An additional three discourse positions are further defined as paragraph-initial, -medial and -final to indicate the initiation, continuation and termination of a speech paragraph. The output of global discourse prosody of a multi-phrase utterance or paragraph can thus be attributed to layers of contributions from discourse levels and positions. At the same time, percentage of contributions by layer cumulatively derives the prediction of output prosody and systematically accounts for the ultimate output [1, 2]. Note that the account also helps clarify why the context of discourse prosody must include both single-unit neighborhood concatenation and between-unit cross-over association.

2.2.2. Tagging perceived emphases by degree

The same speech data are further manually tagged by trained transcribers into a string of emphasis/non-emphasis tokens (ETs) for 4 degrees of perceived strength of prominence defined as follows:

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

3. METHODOLOGY

Patterns of the distribution of ETs can be derived by each discourse unit the PPh. The same types of patterns are then merged into a unique type and calculated for respective frequency by type. Cumulative frequency distribution (CDF) is adopted [10] and defined below.

$$F_a(X)=P(a\leq X)$$

where the right side of the equation represents the probability that the unique pattern a takes on a value less than or equal to X.

In order to examine whether the PPh length bears any relationship with high frequency ET patterns, averaged PPh length by SYL number is calculated by each unique type of emphasis pattern of PPh. The same rationale is also applied to examine whether the most frequent ET patterns bear any relationship with discourse positions defined by 4 types S, -I, -M and -F: (1) when a paragraph contains one single PPh, discourse position is coded as 'S'. (2) When a paragraph contains more than 2 PPhs, the first (initial) and last (final) PPh are coded as 'I' and 'F' while the other in-between PPhs are coded as 'M' (medial).

4. COMPARISONS OF PATTERNS OF PERCEIVED EMPHASIS

4.1 Distribution of discourse units and emphasis tokens by speech genre

The distribution of the number of emphasis tokens (ETs) by discourse units SYL and PPh and by the 3 speech genres CNA, WB and LEC is calculated and listed is Table1.

Table1. Distribution of discourse units and emphasis tokens by speech genre

Corpora / Discourse units and emphasis tokens	CNA	WB	LEC
# of SYL	11594	7061	7660
# of PPh	1490	745	954
# of ET pattern	37	31	60

4.1.1. Discussion

The results show that unique distribution of emphasis patterns by PPh in spontaneous speech LEC are almost twice the amount than either kind of read speech while little difference is found in read speech (LEC/60 CNA/37 WB/31) suggesting that spontaneous speech features more accentuation than read speech and can be described as prosodically more expressive .

4.2 Distribution of emphasis pattern at the PPh level

We clustered the frequency of emphasis pattern by PPh as a test to see if common patterns can be derived; emphasis pattern is defined as the sequence and status of tagged emphasis. The results showed that nearly 70% of PPhs examined can be accounted for by 6 emphasis patterns. The 6 most frequent cross-genre emphasis patterns are (1) E1, (2) E2 E1, (3) E1 E2 E1, (4) E1 E2, (5) E2 and (6) E2 E1 E2. We then calculated the frequency of each type used by genre and compared the distribution of frequency, as shown in Figure 1.

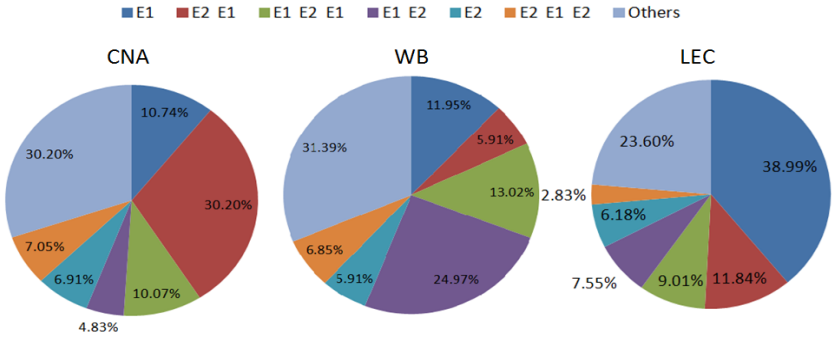

Figure1. The distribution of emphasis patterns (ET type and sequence) by PPh in read speech CAN, WB and spontaneous speech LEC.

4.2.1 Discussion

First of all, the results show genre-independent common components do exist across speech genre while their respective distribution of frequency by each genre is quite different. Secondly, by examining the frequency distribution across genre, the most discriminative patterns are patterns 'E2 E1' 'E1 E2' and 'E1' for CAN, WB and LEC respectively. While spontaneous speech LEC can be discriminated from read speech CNA and WB by the pattern 'E1' (38.99% vs. 10.74% and 11.95%). The distribution of the 2-ET pattern "E2 E1" and "E1 E2" are also highly discriminative across the three genres, namely, 30.2%/5.91%/11.84% and 4.83%/24.9%/7.55% for CNA, WB and LEC, respectively. Moreover, the two genres of read speech CNA and WB can be further be discriminated by 'E2 E1' (30.20% vs. 5.91%) and 'E1 E2' (4.83% vs. 24.97%). The above results demonstrated that genre related prosodic expressiveness is more related to distribution of emphasis pattern than type of emphasis pattern, suggesting that the degree of focus, as perceived emphasis, may be few in type, but the expressive variation in output manifestation is achieved through emphasis allocation instead of type.

Not shown in Figure 1 are the next genre-discriminative patterns that sets read speech apart from spontaneous LEC are more complex patterns 'E2 E1 E2 E1' and 'E1 E2 E1 E2', respectively. The secondary results suggest that emphasis patterns in spontaneous lecturing are in general simpler than read speech, another possible genre feature. However, our previous findings showed that the mean length of PG in LEC is approximately 6 times the length of CNA (653.09 syllables vs. 76.76 syllables) and in faster speaking rate [11]. Viewed with this additional piece of information, we propose that spontaneous lecturing involves well-planned theme and focus of information, and is characterized by packed large amount of information in long paragraphs composed of shorter individual phrases and less complex emphasis patterns at the phrase level. In other words, viewed phrase by phrase, spontaneous lecturing may seem simpler in emphasis pattern. But viewed by paragraph, the overall information allocation may still be more complex. Fourthly, cross-genre comparison by emphasis pattern and PPh length (mean number of syllables per PPh) reveals an interesting result. The mean length of PPh by genres CAN/WB/LEC for "E1" is 3.36/4.12/4.85, for "E2 E1" 6.38/7.20/8.63 and for "E1 E2" 5.36/7.02/6.90, respectively. In other words, the number of emphasis per phrase is positively correlated to phrase length. The PPh that contains only one ET (normal stress only) is under 5 syllables across genre while those containing two ETs are proportionally longer. In addition, PPh in spontaneous speech LEC is slightly longer than read speech. As for more complex emphasis patterns (and not shown due to space), we note that emphasis patterns of three or more ETs are 9 syllables or more in length, indicating that the more complex the emphasis patterns are, the longer the bearing phrase needs to be. We therefore argue that the very short 'E1'-only pattern in LEC and its high frequency distribution may be due to its function as filled pauses, another feature that may be genre-dependent for lecturing. In addition, The "E2 E1" pattern that distinguishes read speech CNA from the other two genres can be regarded as a specific feature of reading prose; while the "E1 E2" pattern that distinguishes simulating weather forecast from prose reading and spontaneous lecturing can be regarded as a specific feature of WB.

The above results show different styles of prosodic expressiveness are exhibited through varied distribution of shared common emphasis patterns.

4.3 Emphasis patterns of PPh by discourse position and speech genres

To observe whether genre specific-ness can be attributed to effects from discourse positions, each of the 6 most frequent common patterns in each genre is examined in relation to discourse features S, I, M and F. The results are listed in Table 2.

Table2. Distribution of emphasis patterns by discourse position S, I, M, F and speech genre CNA, WB, LEC.

Dis position / Emphasis pattern /Speech genre		S	I	M	F
E1	CNA	0.00%	26.88%	51.88%	21.25%
	WB	0.00%	10.11%	76.40%	13.48%
	LEC	0.81%	18.60%	52.02%	28.57%
E2 E1	CNA	3.33%	26.44%	45.33%	24.89%
	WB	0.00%	50.00%	34.09%	15.91%
	LEC	4.42%	13.27%	53.98%	28.32%
E1 E2 E1	CNA	2.67%	16.00%	46.00%	35.33%
	WB	1.03%	14.43%	51.55%	32.99%
	LEC	3.49%	24.42%	41.86%	30.23%
E1 E2	CNA	1.39%	13.89%	59.72%	25.00%
	WB	0.00%	16.13%	69.89%	13.98%
	LEC	0.00%	36.11%	47.22%	16.67%
E2	CNA	0.00%	42.72%	52.43%	4.85%
	WB	0.00%	7.89%	86.84%	5.26%
	LEC	0.00%	20.34%	64.41%	15.25%
E1 E2 E1	CNA	5.71%	24.76%	47.62%	21.90%
	WB	0.00%	15.69%	58.82%	25.49%
	LEC	0.00%	29.63%	55.56%	14.81%

4.3.1 Discussion

The results show the distribution of emphasis allocation by discourse position, genre difference is found. For PPhs at the discourse-initial position, the highest percentage of emphasis pattern is 'E2' for CNA (42.72%), 'E2 E1' for WB (50 %) and 'E1 E2' for LEC (36.11%%), respectively. For PPhs in discourse-medial position, the highest percentage of emphasis pattern is 'E1 E2' for CNA (59.72%), 'E2' for WB (86.84 %) and 'E2' for LEC (64.41%) respectively. For PPhs in discourse-final position, the highest percentage of emphasis pattern is 'E1 E2 E1' for CNA (35.33%), 'E1 E2 E1' in WB (32.99%) and 'E1 E2 E1' for LEC (30.23 %) respectively.

By discourse-Initial/Medial position, 'E2' is found as a frequent pattern. It appears at discourse-Initial position in LEC but -medial position in WB and LEC, and is therefore genre specific. The 'E1 E2 E1', on the other hand, is found in BG-final positions regardless of discourse position and can be regarded as a genre-independent feature.

5. CONCLUSION

From the results presented, we found that the allocation of key information, perceived as emphasis in continuous speech, may appear to be random and highly varied on the surface. However, the above results demonstrate that their patterns are relatively few while allocation is in fact systematic. In other words, more expressive prosody output is fine-tuned by highlighting key information across the speech flow while genre or style related prosodic expressiveness is also organized instead of random. We found that both genre-independent emphasis patterns and genre-specific features collectively deliver more prosodic expressiveness in addition to discourse association. The genre related stylistic differences are achieved by varied distributions of the same patterns, and the most discriminative patterns are found. Genre-specific and genre-independent patterns by discourse position are found as well, indicating that while discourse association is maintained, prosodic expressiveness can be different. In addition, patterns of emphasis sequencing are positively correlated to length of the emphasis bearing unit. More complex emphasis patterns require longer bearing unit. Lastly but not surprisingly, no two successful but identical ET are found regardless of the complexity of emphasis pattern because by default no contrast can be achieved if the emphases are of the same degree. Future research includes more refined analysis in relation to information structure for the quantitative aspects of continuous speech prosody, and qualitative exploration of the other contributing factors to prosodic expressions.

6. REFERENCES

[1] Tseng, C., 2010. "Beyond Sentence Prosody". Interspeech. Makuhari, Japan.

[2] 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.

[3] 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.

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

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

[6] Tseng, C., and Su, Z., 2007. "From One Base Form to Multiple Output Styles-- Predicting Stylistic Dynamics of Discourse Prosody". Interspeech 2007 Eurospeech 110-113. Antwerp, Belgium.

[7] 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, 2005.

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

[9] 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.

[10] Hogg, R.V., 2001. Probability and statistical inference (6th ed.). Prentice Hall, NJ: Upper Saddle River.

[11] Tseng, C., and Su, Z., 2008. "Spontaneous Mandarin Speech Prosody—the NTU DSP Lecture Corpus". Oriental COCOSDA 2008 171-174. Kyoto, Japan