Does Semantic Stress Have Effect on Duration and Pitch Patterns of Prosodic Words in Presenters' Speech?

Yu ZOU, Wei HE, Min HOU, and Yonglin TENG
Broadcast Media Language Branch, National Language Resources Monitoring & Research Center
Communication University of China, Beijing
{zouiy, hewei, houmin, tengyonglin}@cuc.edu.cn

Abstract—**Stress is a main prosodic feature; it becomes more and more important in speech recognition, speech synthesis and understanding, especially in the spontaneous speech. This paper is intended to analyze whether the semantic stress has effect on pitch and durational pattern of prosodic word in broadcast presenters' speech. The research results shows that the stresses will not make a fundamental difference to change the durational pattern most of which still keep a long-tail duration pattern within prosodic words. Whereas, stresses can influence the duration of some particular syllables, for instance, in some four-syllable prosodic words, the durational pattern changes to a long-head pattern if its first two syllables are accented. On pitch pattern, there are different prominences because the different syllable is accented within prosodic word. Thus the pitch feature of stressed prosodic words manifests more complex than that of the non-stressed ones. To some extent, we can suggest that the pitch pattern of stressed prosodic words has been changed by semantic stress.**

Keywords-semantic stress; duration pattern; pitch pattern; prosodic word; broadcast presenters' speech

I. INTRODUCTION

Stress is an important feature of prosody, which plays an important role of the semantic expression in communication each other, and also plays a very important role of speech processing for natural language understanding, speech recognition and speech synthesis etc. In linguistics, stress is the relative emphasis that may be given to certain syllables in a word, or to certain words in a phrase or sentence. The term is also used for similar patterns of phonetic prominence inside syllables. The ways stress manifests itself in the speech stream are highly language-dependent.

In English, stress is most dramatically realized on focused or accented words. For instance, consider the dialogue "Is it brunch tomorrow?" "No, it's *dinner* tomorrow." In it, the stress-related acoustic differences between the syllables of "tomorrow" would be small compared to the differences between the syllables of "*dinner*", the emphasized word. In these emphasized words, stressed syllables such as "*din*" in "*dinner*" are louder and longer [1] [2]. On the other hand, stressed and accented syllables are produced with additional lengthening compared with unstressed syllables [3].

In Chinese, Lin also argued that the stress is part of intonation, because the different manifestations of F0 scale, F0 range and duration in each syllable of utterance are conditioned largely by stress [4]. Some other researchers investigated the influence of focus on durational patterns of five-syllable words with various positions and different tones based on experimental speech materials, their research results showed that although focus induces significant lengthening of the focused constitutes, the internal durational adjustment of each focused syllable is by no means symmetric and the magnitude of such lengthening is determined by the metrical structure of the focused constituents [5].

However, our earlier investigations of broadcast presenters' speech revealed that there is some difference of prosodic structure/feature compare to the experimental speech [6] [7]. In that way, does semantic stress have effect on duration and pitch patterns of prosodic words in presenters' speech? And which changes does it happen if the semantic stress can affect the prosodic pattern? This paper is intended to analyze whether the semantic stress has effect on pitch and durational pattern of prosodic word in broadcast presenters' speech.

The paper is organized as follows: Section 2 of this paper describes speech data. The perception experiment is designed and carried out in section 3. Section 4 is dedicated to analysis and discussion of the duration and pitch pattern within stressed prosodic words. Finally, some conclusions and outlines of our future work are given in section 5.

II. DATA

A large-scale broadcast language corpus is available at Communication University of China (CUC). This corpus consists of two closely related to the sub-corpus, text and speech. It includes different periods of radio and television audio and video database, in particular various periods of the award TV or radio programs.

In our previous study, four speaking styles was been defined including news announcing, oral interpretation, explanation and talking based on state-of-the-art classifier in broadcast speech corpus. The presenter's speech is a typical talking style. About 70 minutes, 13787 syllable conversation corpora were selected for this study. The selected speech data contains four presenters (two male and two female) is fairly representative of the broadcast speech corpus.

All the speech data sampling rate is 22 kHz, resolution is 16bit, mono, the file saved as PCM wav format. The annotation and analysis tool is Praat. Data preprocessing use MS Excel 2007, statistical analysis tool is SPSS 13.0. The fundamental

This paper is supported by the Department of Science and Technology at Ministry of Education (No. 107118), and the "211" Key Projects of Communication University of China (No. 21103010105, 21103010106).

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

frequency was normalized by semitones, the normalization formula is $ST=12*\log(f/f_{ref})/\log 2$, Where f is one fundamental frequency which will be normalized, f_{ref} is the reference frequency. In this study, the male's reference frequency is 55Hz, female is 100Hz.

III. PERCEPTIONAL EXPERIMENT

A. Experimental Procedure

Firstly, we designed a series perception experiments to determine the prosodic boundary of prosodic word, prosodic phrase, intonational phrase and intonational group. The selected speech data which contain 13787 syllables, 68 wav files were played to eleven listeners of native Chinese, five male and six female. Most of them are doctoral or master's candidate, there were also some teachers and related researchers.

Secondly, the 68 wav files were released again to the listeners who were asked to identify the semantic stress in each sentence.

Every wav file was released 3-4 times with an interval of 20 seconds. After one hour of continuous work, a 20-minute break was given.

B. Experimental Results and Analysis

After the experiment, only the results with at least a 90% agreement rate were considered for analysis. Whereas, we find many stress, which were determined by 90% agreement rate, are marked as a whole by the listeners, such as *zhu2zi0* (竹子 bamboo), *jin1zi0* (金子 gold), *fu4qin0* (父亲 father), *gan3ren2 de0* (感人的 affecting), *feng1mian4 shang0* (封面上 on the book cover) and *zheng1da4 yan3jing0* (睁大眼睛 widely open eyes) and so on. These are invalid because the last syllable is neutral tone, it cannot be accented in Mandarin Chinese. At the post-processing step, we corrected them by hand.

As a result of the post-processing step, we put the valid semantic stresses in prosodic words to consider their prosodic feature. If there is one, two, three or four syllables were accented within a prosodic word, we look it as a stressed prosodic word, and compare the durational and pitch pattern between them and the unstressed ones in our work.

IV. ANALYSIS AND DISCUSSION

To facilitate analysis and discussion, four kinds of prosodic words, which are one-syllable, two-syllable, three-syllable and four-syllable respectively, were classified by the number of syllables. Their distribution is shown in Table 1. However, we do not consider the one-syllable prosodic word in this paper.

TABLE I. THE DISTRIBUTION OF PROSODIC WORDS IN OUR DATA.

One-syllable PW	Two-syllable PW	Three-syllable PW	Four-syllable PW
45	144	179	298

A. Duration Patterns of Prosodic Word

According to the distribution of the data, we analyze the comparative duration data of prosodic words on normal and stressed that they locate at the middle and tail of sentence. Fig. 1, 2 and 3 show the comparative duration data of two-syllable, three-syllable and four-syllable prosodic words on normal and stressed, respectively.

There is only one example of stressed prosodic words at the head of sentence in our data. It is very sparse. We thus consider the prosodic words on normal and stressed which locate the middle and tail of the sentence.

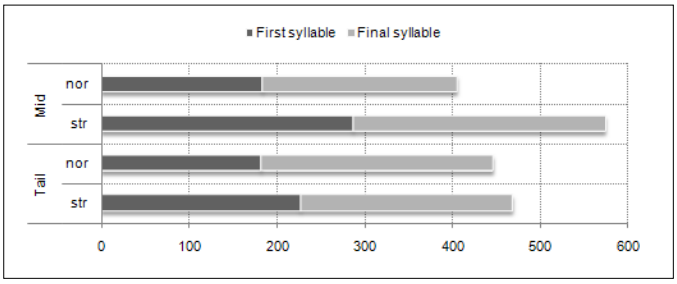

Figure 1. The comparative duration data of two-syllable prosodic words on normal and stressed (milliseconds). ("nor" is the abbreviation of normal, "str" is the abbreviation of stressed)

In two-syllable prosodic word, that the two syllables are all stressed is perceived by the listeners in our perceptional experiment. By observing the duration data of stressed prosodic words in Fig. 1, we find that the mean of duration of each syllable that lie in the middle of the sentence is roughly equal, that is 286.4 and 286.9 milliseconds respectively. Furthermore, at the tail of sentence, the mean of duration of last syllable is slightly longer than that of first one.

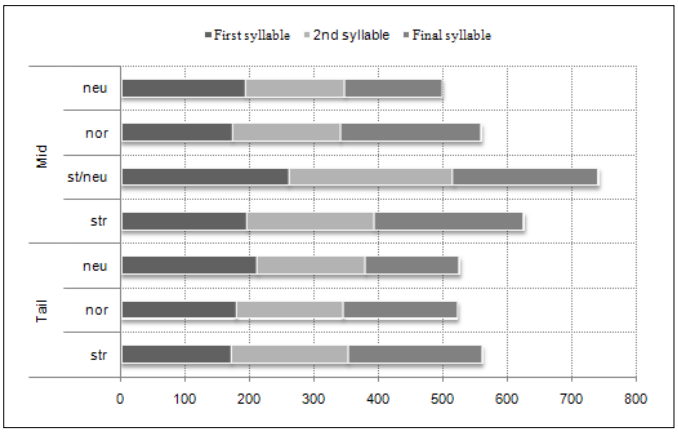

Figure 2. The comparative duration data of three-syllable prosodic words on neutral, normal and stressed (milliseconds). ("neu" is the abbreviation of neutral, it means that the last syllable is neutral tone in the three-syllable prosodic word; "st/neu" means that it is a stressed prosodic word which the last syllable is neutral one)

There are two kinds of stressed three-syllable prosodic words based on the results of experiment: if there is no neutral tone, the three syllables are all accented; otherwise, the first two syllables are prominence if the last syllable is neutral tone. The latter does not appear at the tail of sentence in our data.

According to the duration data of stressed three-syllable prosodic words in Fig. 2, except the neutral prosodic words (including one stressed prosodic words which the final syllable is neutral tone), the stressed and normal ones both have a long-tail durational pattern. For instance, the mean of duration of each syllable in the stressed one is 171.1, 181.6 and 208.9 milliseconds respectively; similarly, the normal one is 178.4, 166.5 and 178.2 milliseconds respectively.

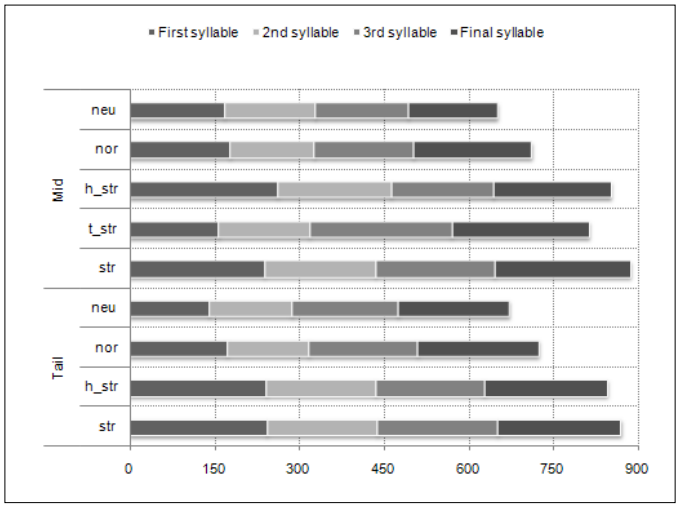

Figure 3. The comparative duration data of four-syllable prosodic words on neutral, normal and stressed (milliseconds). ("h_str" means that the first two syllables are stressed; "t_str" means that the last two syllables are stressed; "str" means that all four syllables are stressed)

In the four-syllable prosodic word, there are three kinds of prominence: 1) all syllables, 2) the first two syllables and 3) the last two syllables are accented. Fig. 3 shows the details. According the stressed prosodic words which locate the middle of sentence, we know that mean of duration of the first and final syllable is almost equal, that of the second and third one is relative shortening, if all syllables are stressed. For example, the mean of duration of each syllable is 237.6, 196.3, 210.1 and 239.7 milliseconds, respectively. The mean of duration of the final syllable is the longest than that of others if the first two syllables are prominence (e.g. its mean of duration are 260.7, 200.9, 180.5 and 207.2 milliseconds successively); on the contrary, that of the first one is shortest if the last two syllables are stressed (e.g. its mean of duration are 157.2, 160.1, 251.1 and 242.3 milliseconds successively). Therefore, the extended range of the mean of duration in the whole stressed four-syllable prosodic words are about 310 milliseconds. There are almost equal results will be concluded according to the data that lie in the tail of sentence.

Comparison the data of normal prosodic words from above figures, we can find that there is little change of global durational pattern on prosodic words because of the prominence by semantic stress. The global durational pattern of the majority of prosodic words still is long-tail pattern. But the local durational pattern of the part of prosodic words is a certain impact. Furthermore, the mean of total duration of normal prosodic words is between 330.51 and 722.01 milliseconds, it has about 391.5 milliseconds extended range, and that of the stressed one thus is between 467.8 and 883.7 milliseconds with about 416 milliseconds extended range. So the prominence of semantic stress cause lengthening of the accented prosodic word, and the lengthened range also become longer. For instance, the mean of total duration in the stressed prosodic words have about 169, 243 and 174 milliseconds longer than that of normal ones in two-syllable, three-syllable and four-syllable respectively.

B. Pitch Feature of Prosodic Word

As for the pitch feature of prosodic words, there are many previous investigations have been reported. For instance, the extension of the bottom of range within utterance reflects the rhythm structure, and the rising about the top of range is related to the semantic emphases. Therefore, the prominent feature of stress is the prominence of top pitch value [8]. Wang et al [9] argued that the pitch movement of stressed word is on basis of top-and bottom-line declination intonation pattern. The rising of high point of the pitch is the main cue to stressed syllable while the movement of low point of the pitch is not that much and is limited by the intonational bottom-line declination.

Some other researchers also reported that the distribution of focus-related accents within sentences and the distribution of semantic accents within phrases in Chinese through 300 natural utterances. Their experimental results show that the semantic accent tends to be post-posed in S-P and V-O structures, while pre-posed in adjunct-head structures [10]. However, is there any effect on the pitch feature in prosodic word? The detailed data of prosodic words on stressed and normal are shown in Table 2, 3, 4, 5 and 6, Fig. 4, 5 and 6, respectively.

TABLE II. THE PITCH DATA OF TWO-SYLLABLE PROSODIC WORDS WHICH IS ACCENTED BY SEMANTIC STRESS (SEMITONES).

Location	**Mid of sentence**				**Tail of sentence**			
	First syllable		***Final syllable***		***First syllable***		***Final syllable***	
	Str.	Nor.	Str.	Nor.	Str.	Nor.	Str.	Nor.
top[a]	22	16.2	18.1	14.9	15.6	14.5	10.1	10.1
SD[b]	4.1	4.1	3.6	4.6	0.3	1.2	0.1	1.1
bottom[c]	10.3	12.1	7.7	9.4	10.5	--	6.6	--
SD	4.7	4.3	6.3	3.4	6.6	--	0.6	--
range	11.7	4.1	10.4	5.5	5.1	--	3.5	--

a. "top" is the mean of the highest pitch value at the first tone and the fourth tone.

b. "SD" is the abbreviation of standard deviation.

c. "bottom" is the mean of the lowest pitch value at the third tone and the fourth tone.

TABLE III. THE PITCH DATA OF THE STRESSED THREE-SYLLABLE PROSODIC WORDS, EXCEPT THE "ST/NEU" (SEMITONES).

Location	**Mid of sentence**			**Tail of sentence**		
	1st syl.	***2nd syl.***	***3rd syl.***	***1st syl.***	***2nd syl.***	***3rd syl.***
top	12.5	16.3	13.8	13.1	17.3	14.4
SD	1.3	5.0	5.2	2.1	2.6	5.9
bottom	7.6	3.9	8.3	4.1	9.2	7.0
SD	3.6	3.4	4.9	1.3	3.6	0.3
range	4.9	12.4	5.5	9.0	8.1	7.4

TABLE IV. THE PITCH DATA OF NORMAL THREE-SYLLABLE PROSODIC WORDS (SEMITONES).

Location	Mid of sentence			Tail of sentence		
	1st syl.	*2nd syl.*	*3rd syl.*	*1st syl.*	*2nd syl.*	*3rd syl.*
top	17.0	16.4	14.5	16.5	12.2	10.1
SD	4.1	4.7	3.7	5.4	6.6	9.6
bottom	12.2	10.9	8.4	11.4	5.8	8.1
SD	3.7	3.0	4.3	4.9	5.5	7.7
range	4.8	5.5	6.1	5.1	6.4	2.0

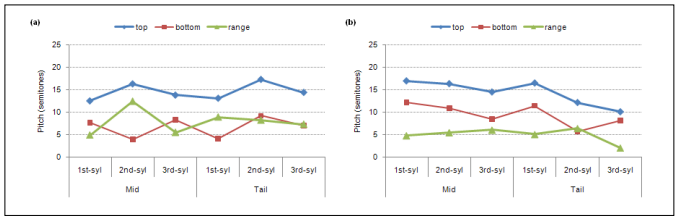

Figure 4. The comparative pitch data of three-syllable prosodic words at middle and tail of the sentence on (a) stressed and (b) normal one (semitones).

In our data, we did not get the bottom pitch value of normal two-syllable prosodic word at tail of the sentence. From Table 2, the top and bottom pitch value of first syllable are higher than that of final syllable in the two-syllable stressed prosodic words, and the range of the head is wider than that of the tail. This illustrates that the first syllable is more prominence.

According to Table 3 or Fig. 4(a), we can know that the top of 2nd syllable is highest than that of others in stressed three-syllable prosodic words which locate the middle of sentence, and its bottom also is lowest, it has a widest pitch range accordingly. At the tail of sentence, the top and bottom of the 2nd syllable are both raised, the pitch range just has little change, for instance, their range are 9.0, 8.1 and 7.4 semitones successively. Comparison Table 4 or Fig. 4(b), the top-and bottom-line both have apparent declination from head to tail within normal prosodic word, and the pitch range does not clearly fluctuates except that of the final syllable at the tail of sentence. To some extent, we can indicate that the pitch pattern of three-syllable prosodic words has been changed by semantic stress.

TABLE V. THE PITCH DATA OF NORMAL FOUR-SYLLABLE PROSODIC WORDS (SEMITONES).

Location	Mid of sentence				Tail of sentence			
	1st Syl.	*2nd Syl.*	*3rd Syl.*	*4th Syl.*	*1st Syl.*	*2nd Syl.*	*3rd Syl.*	*4th Syl.*
top	17.9	15.6	15.8	14.2	15.0	13.4	12.6	9.4
SD	4.1	4.7	4.1	4.7	3.9	6.3	5.1	4.8
N	59	51	54	51	30	22	29	31
bottom	11.9	11.5	10.3	8.7	9.3	5.6	7.0	6.9
SD	5.0	4.0	4.6	6.2	6.3	6.7	5.8	5.4
N	51	54	42	61	27	19	26	27
range	6.0	4.1	5.5	5.5	5.7	7.8	5.6	2.5

TABLE VI. THE PITCH DATA OF FOUR-SYLLABLE PROSODIC WORDS WHICH ALL SYLLABLES ARE STRESSED (SEMITONES).

Location	Mid of sentence				Tail of sentence			
	1st Syl.	*2nd Syl.*	*3rd Syl.*	*4th Syl.*	*1st Syl.*	*2nd Syl.*	*3rd Syl.*	*4th Syl.*
top	18.8	17.7	19.3	17.4	21.3	18.3	15.9	10.3
SD	0.4	5.0	3.5	1.4	4.4	5.4	5.4	4.6
bottom	11.2	10.1	0.9	2.0	6.3	7.5	5.5	3.7
SD	4.1	6.9	4.1	3.7	4.0	4.5	2.7	4.7
range	7.6	7.6	18.4	15.4	15.0	10.8	10.4	6.6

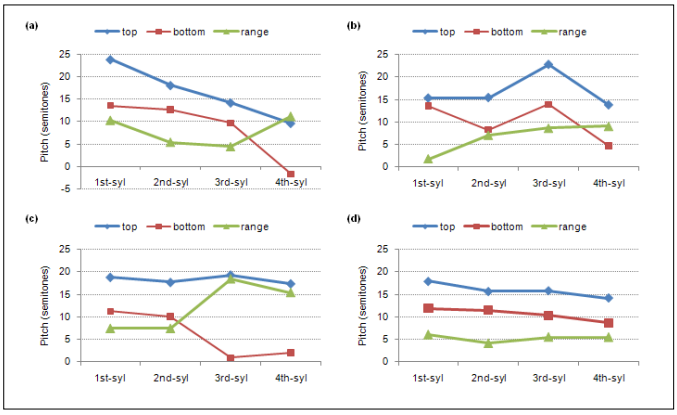

Figure 5. The comparative pitch data of four-syllable prosodic words at middle of the sentence on (a) the first two syllables, (b) the last two syllables, (c) all syllables are stressed, and (d) all syllable are normal (semitones).

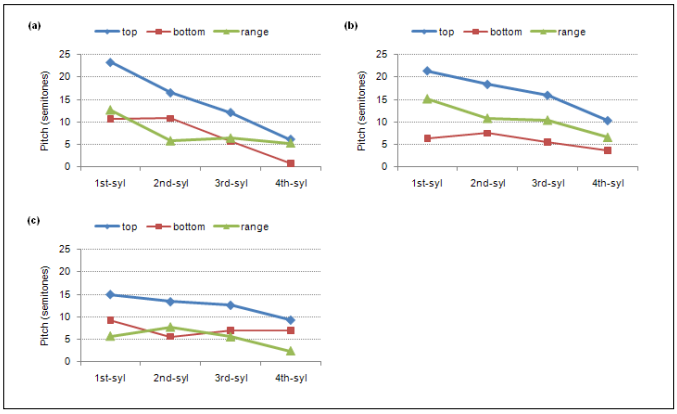

Figure 6. The comparative pitch data of four-syllable prosodic words at tail of the sentence on (a) the first two syllables, (b) all syllables are stressed, and (c) all syllables are normal (semitones).

From Fig. 5(a), we find that the top-and bottom-line are declination successively, and there is a widely drop of top and bottom pitch value at the final syllable. Especially the bottom, it has about 10 semitones sudden drop. The pitch range has about 7 semitones room for fluctuating within the four-syllable prosodic word.

According to Fig. 5(b), there is a widely rising (about 7 semitones) of the top-line at the third syllable, and a sudden drop (about 9 semitones) at the final syllable. The bottom-line

has a similar change at the back of the four-syllable prosodic word.

By observing Table 6 and Fig. 5(c), if all syllables are accented by perception, the top-line has no clearly fluctuation, and the bottom-line has an abrupt fall from the second syllable (10.1 semitones) to the third one (0.9 semitones). The pitch range of the tail broadens wider than that of the head (from 7.6 to 18.4 semitones).

Comparison Table 5 or Fig. 5(d), there is a smooth declination from head to tail within normal prosodic word, and the pitch range does not clearly fluctuating. This is the difference between the normal and the stressed four-syllable prosodic words at middle of the sentence.

The detailed pitch data of four-syllable prosodic words at tail of the sentence is shown in Fig. 6 and Table 5 and 6. According to (a), (b) and (c) part of the figure, we can indicate that the top-line has little fluctuation besides it hold a clearly declination. The bottom-line just has a little fluctuation at the second syllable. The pitch range thus is gradually narrowing from head to tail. The range of the stressed prosodic words is about 3 semitones more than that of the normal one for change. Just in this situation, we can say that the stress has a little effect on the pitch pattern at tail of the sentence.

V. CONCLUSIONS

Stress (especially semantic stress) becomes more and more important in speech recognition, synthesis and understanding, especially in the spontaneous speech. However, whether does it affect the duration and pitch pattern within prosodic words in broadcast presenters' speech? This paper investigated the prosodic feature of two-, three-and four-syllable prosodic words. According to the results, we can draw some conclusions as follow:

Firstly, there is not a fundamental difference of durational pattern between the stressed prosodic words and the non-stressed ones. The majority of accented prosodic words still keep a long-tail pattern. Whereas, stresses can influence the duration of some particular syllables, for instance, in some four-syllable prosodic words, the durational pattern changes to a long-head pattern if its first two syllables are accented because of prominence. Comparison the non-stressed prosodic words, the total duration of stressed ones is longer.

Secondly, there are different prominences because the different syllable is accented within prosodic word. Thus the pitch feature of stressed prosodic words manifests more complex than that of the non-stressed ones. For example, the top-line has larger fluctuating, obvious declination within the stressed prosodic words; the top of the second syllable is about 3-4 semitones higher than that of the first and last syllable in stressed three-syllable one. Another example, there is a widely rising of the top-line at the third syllable, and a sudden drop at the final syllable if the last two syllables will been accented within the four-syllable prosodic words, and its bottom-line has a similar change.

Future research will include treatment of whether the semantic stress has effect upon duration and pitch of prosodic phrase or intonational phrase, ideally comparing the different affect between announcing style and talk style in broadcast speech. Additionally, we would like to tackle the correlation between semantic/syntactic structure and prosodic structure within prosodic word or prosodic phrase.

ACKNOWLEDGMENTS

We would like to thank Dr. Ziyu Xiong for his PRAAT script, and the anonymous reviewers for their insightful comments.

REFERENCES

[1] M. E. Beckman, Stress and Non-Stress Accent. Dordrecht: Foris, 1986.

[2] R. Silipo and S. Greenberg, "Automatic Transcription of Prosodic Stress for Spontaneous English Discourse", in Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS XIV), San Francisco, CA, August 1999, pp. 2351-2354.

[3] M. E. Beckman and J. Edwards, "Articulatory evidence for differentiating stress categories", in P. Keating (eds.), Phortological Structure and Phonetic Form: Papers in Laboratory Phonology III. Cambridge: Cambridge University Press, 1994, pp. 7-33.

[4] M. C. Lin, "Prosodic structure and lines of F0 top and bottom of utterance in Chinese", Journal of Contemporary Linguistics, 4(4), 2002, pp. 254-265.

[5] Y. Jia, A. J. Li, Z.Y. Xiong and Y. Y. Chen, "Effects of Focus Upon Durational Patterns of Five-syllable Words in Standard Chinese", in Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS XVI), Saarbrücken, 6-10 August 2007, pp. 1181-1184.

[6] Y. Zou, "A Formal Study on Prosody of Presenter's Spoken Language Based on Broadcast Speech Corpus", PhD thesis, Communication University of China, 2007.

[7] Y. Zou, W. He, Y. Q. Zhang, M. Hou, and W. B. Zhu, "A Special Prosodic Phrasing in Broadcasting News Programs", in Computational Sciences and Optimization: Theory, Simulation and Experiment (Vol. 2), the IEEE Computer Society, USA, 2009, pp. 406-408.

[8] J. Shen, "Intonation structure and intonation types of Chinese", Dialect [*Fangyan*], 4, 1994, pp. 221-228.

[9] B. Wang, S. N. Lv and Y. F. Yang, "The Pitch Movement of Stressed Syllable in Chinese Sentences", Acta Acustica, 27(3), 2002, pp. 234-240.

[10] Y. J. Wang, M. Chu and L. He, "An experimental study on the distribution of the focus-related and semantic accent in Chinese", Chinese Teaching in the World [*Shijie Hanyu Jiaoxue*], 76(2), 2006, pp. 86-98.