Prosodic Boundary Perception in Spontaneous Speech of Standard Chinese

Aijun Li

The Institute of Linguistics, Chinese Academy of Social Sciences, China

Email: liaj@linguistics.cass.net.cn

ABSTRACT

A perception experiment was carried out to determine the perceived prosodic boundary strength and the hierarchies of the perceived prosodic boundaries for spontaneous Chinese. The results indicate that perceived prosodic boundaries are highly correlated between intelligible and unintelligible (delexical) utterances, showing that the listeners may judge prosodic boundaries with no reference to semantic, syntactic and lexical information and that current method of prosodic annotation is feasible and scientific. The prosodic boundaries annotated by experts are highly correlated to those done by native listeners through clustering the perceived prosodic boundaries into the same levels. So untrained listeners can give reliable annotation and a learning sample set for prosodic break annotation can be produced from the perceptual classification. The perceived breaks forms a continuum, which can produce numerous boundary levels theoretically and a blurry part can exist between each two adjacent levels.

1. INTRODUCTION

The investigation to prosodic boundary is useful for both speech synthesis and speech recognition. Various analyses on prosodic boundaries have been carried out including perceptual analysis[3,4,8,10,11,13], acoustic analysis [5,14,18,20] and automatically segmentation[9]. Jan Roelof de Pijper and Angelien A. Sanderman made an investigation on perceived boundary strength (PBS) and the phonetic cues for Dutch and found that "the untrained listeners can give reliable and usable judgment of PBS and that is true even if the lexical contents of the utterances are made unrecognizable thus blocking access to lexical, syntactic, and semantic information." [8] But is it true for a tone language like Chinese?

As we have known that syntactic boundary and the related prosodic boundary are not isomorphic and many contributions have described the relationship between these two kinds of boundaries from different aspects of phonetics, phonology, syntax, and pragmatics [3,6,7,16,20]. C-ToBI is a ToBI-like transcription system developed for Chinese prosodic annotation [2]. The prosodic boundary levels or prosodic structures of Chinese are based on the pre-investigations from two perspectives[5,7]: phonological-syntactic, and phonological-phonetic perspectives. Based on these studies, we define the hierarchically organized prosodic structure from small to large constituents as syllable, prosodic word (PW), minor phrase (MIP), major phrase (MAP) and intonation group (IG). But by now, no perception experiment has been made to define the correlation between the PBS and the prosodic levels. Especially for spontaneous speech whose prosodic boundaries distribute and are realized differently from the read speech [4,19], it will be very useful and important to study how to establish a feasible and scientific prosodic boundary annotation criteria based on PBS. Therefore the main issues to be addressed in this paper are the following:
(1) The annotation reliability: Are the annotation results produced by experts the same as those done by non-expert or native listeners?
(2) The resolution of the problem: Is correct prosodic annotation for the meaningful utterances, also correct under 'semantic distraction.'
(3) How many prosodic levels can we define and how to define the criteria for the hierarchical annotation? Is it possible to make the boundary hierarchies summarized from the PBS of all kinds the criteria for annotation?
(4) The acoustic features of the prosodic boundaries of spontaneous Chinese speech and the differences between the read and the spontaneous speech. (The results are presented in detail in another paper also in this proceeding entitled "Cues of Prosodic Boundaries in Chinese Spontaneous Speech.")

2. PERCEPTUAL EXPERIMENT

2.1 SPEECH MATERIAL

As indicated in the introduction, to find out the influence of the semantic, syntactic and lexical to the prosodic boundary perception, we prepare two stimulus sets: lexical set and delexical set.

The lexical set includes 24 utterances, all female speech, selected from spontaneous speech corpus CADCC by phoneticians, covering as many prosodic boundaries as possible and excluding "abnormal" boundaries in spontaneous speech like repetitions, disfluency, etc.[1] In order to make the perception experiment be easily finished for the delexical set, we have to make sure that one utterance has one related boundary as a target boundary to be judged and this boundary must be the strongest one and all other potential boundaries are far more weaker than this. The text of sample utterances is listed in table 1.

The delexical set can be achieved from the lexical set through an acoustic process which can filter the lexical,

ISBN 1-876346-48-5

syntactic and semantic information from the utterance and retain duration and F0 information. There are several ways to get this delexicalization. Lehiste & wang used a spectrum inversion technique[11]. Lehiste used a band-pass filter when making a perception test for sentence and paragraph[10]. Jan Roelof de Pijper and Angelien A. Sanderman, however, used a more complicated LPC technique to make all the vowels a schwalike quality. We adopt a low-pass filter with cut-off frequency 500Hz to get the delexical set. All filtered 24 utterances retain the duration, F0 and the first harmonic frequency and relative amplitude of the corresponding lexical ones.

utterances (Examples)	meaning
1-a1 我学不了|了我觉得	I can't learn I feel.
12-a2 我觉得|就只能是	I think it must be.
14-a3 声学还有个计算机|这方面	Acoustics and computer as well.

Table 1: The text of utterances. "|" for the target boundary

2.2 PERCEPTION ASSIGNMENT

(1) Perceptual boundary strength assignment

There are various methods to assign a score to a PBS. Petra Hansson, for an instance, used Visual Analogue Scale (VAS) to make the listeners mark the PBS on a 100mm line [13]. Jan employed 10 point-scale to score PBS[8]. We use an ABX test method to score every stimulus pair and sum up the points of each stimulus to score the PBS. Take stimuli AB, for an instance, if A's PBS > B's PBS, then A gets 2 points, and B 0 point; if A's PBS < B's PBS, then B gets 2 points, and A 0 point; if they are equal, then A and B get 1 point respectively.

A software with friendly interface has been made to score PBS of each stimulus pair and let the listener to hear the stimuli pair as many times as they want. The PBS scores for some stimulus can be calculated by summing up all the points this stimulus gets by the software automatically.

(2) Perception on Delexical set

There were 10 naive Chinese listeners taking part in our experiment, 5 male and 5 female with no phonetic knowledge. All of them have normal hearing. The same high quality earphones were used for all the listeners during the test. There are 24*32/2=276 stimulus pairs for delecxical set which were played to the listeners randomly by the software. Before listening, the listeners were taught to use the software and told to hear the strongest boundary of each utterance and give judgment on the PBS of each stimulus pair. It took about 2 and half hours for them to finish the task.

(3) Perception on lexical set

One week later, the same ten listeners were invited to join the experiment on the lexical set. They used the same software to score PBS of all 276 intelligible stimulus pairs. It took about 2 hour for them to finish the task.

2.3 STATISTICAL ANALYSIS AND RESULTS

(1) PBS on lexical and delexical sets

Fig 1 and fig 2 show the PBS scores for delexical and lexical sets respectively. X axes stands for the 24 stimuli and Y for the PBS. Ten thin lines are for ten listeners. The bold black curve is the average PBS. The stimuli are drawn in ascending order of the average PBS. The higher the score is, the stronger the PBS is and the higher level the boundary obtains, and vice versa.

We try to answer the question here if 'objectively' perceived boundary strength of the lexical utterances is the same as the 'subjectively' perceived boundary strength of the delexical utterances. In another words, if the annotation criteria for prosodic boundary annotation in C-ToBI is feasible and scientific.

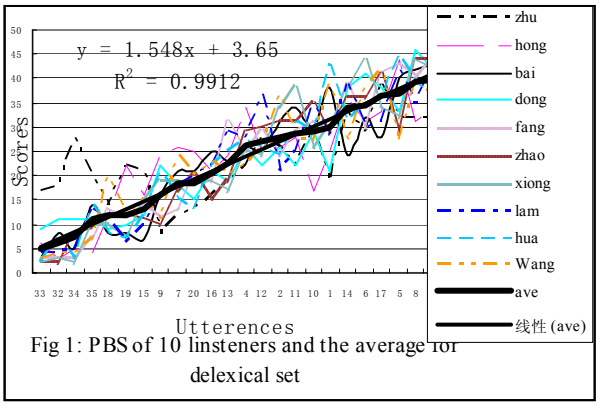

Fig 1: PBS of 10 linsteners and the average for delexical set

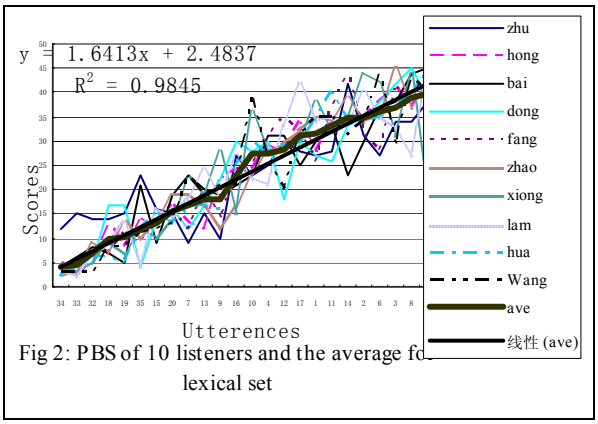

Fig 2: PBS of 10 listeners and the average fo lexical set

(2) Correlation analysis

First of all, we need to make a correlation analysis for ten listeners to get their PBS agreement in lexical and delexical conditions respectively. The results are shown in table 2. The person correlative coefficients reveal the listeners agree very well in their perceptual judgments (correlation is significant at 0.01 level) in both conditions except listeners Zhu-Hong and zhu-Bai in delexical set (significant at 0.05 level).

ANOVA analysis also shows that no significant difference exist among 10 listeners（F=0， df=9, P=0.01) for both lexical and delexical sets; and significant difference exists among 24 utterances (delexical: F=44.72 ; lexical: 58.42, df=3, p=0.01). So the average PBS can be used to predict

the boundary strength and untrained listeners can give reliable and usable annotation.

How about the agreement for the same listener between lexical and delexical conditions? A Person correlative analysis (2 tailed) for each speaker was made between two conditions as shown in table 3 and figure 3. Table 3 and figure 3 show very high correlative coefficients for each listener in two conditions except that Zhu's is a slight lower (r=0.733), but all the correlations are significant at 0.01 level. It reveals that the same listener agrees very well between lexical and delexical utterances and the perceptual judgments for prosodic boundaries are reliable even when the lexical, syntactic and semantic information is blocked.

It is highly correlated between any two subjects in each condition and for the same subject between two conditions, which shows that it has no differences to make prosodic boundary annotation between intelligible and unintelligible phrases. In other words, the listeners may judge prosodic boundaries with no reference to semantic, syntactic and lexical information, which proves that current method of prosodic annotation is feasible and scientific.

	zhu	hong	bai	dong	fang	zhao	xiong	lam	hua	Wang
zhu	!	0.852	0.757	0.776	0.908	0.859	0.633	0.767	0.859	0.713
hong	**0.444***	!	0.912	0.923	0.931	0.942	0.815	0.896	0.954	0.895
bai	**0.499***	0.777	!	0.820	0.861	0.893	0.780	0.795	0.926	0.882
dong	0.709	0.771	0.825	!	0.862	0.875	0.790	0.824	0.882	0.878
fang	0.696	0.849	0.878	0.873	!	0.952	0.744	0.888	0.943	0.828
zhao	0.727	0.782	0.881	0.907	0.930	!	0.804	0.893	0.946	0.871
xiong	0.586	0.817	0.903	0.915	0.883	0.937	!	0.759	0.837	0.901
lam	0.615	0.812	0.880	0.847	0.951	0.889	0.845	!	0.903	0.846
hua	0.567	0.808	0.939	0.845	0.912	0.864	0.893	0.900	!	0.937
wang	0.548	0.748	0.868	0.805	0.851	0.913	0.904	0.776	0.847	!

Table 2: Correlative coefficients between any two listeners (up right angle for lexical set and low left angle for delexical set),* for correlation is significant at 0.05 level, others 0.01 level.

Subject	R^2	Person r
Zhu	0.5368	0.733
Hong	0.7043	0.839
Bai	0.8914	0.944
Dong	0.8022	0.896
Fang	0.8598	0.927
Zhao	0.8298	0.911
Xiong	0.8053	0.897
Lam	0.6608	0.813
Hua	0.8889	0.943
Wang	0.7445	0.8628

Table 3: Correlative coefficients for the same listeners between two conditions. (significant at 0.01 level)

(3) Clustering

From figure 1 and 2, we can see that the curves of average PBS scores are predicted as two linear lines (R^2=0.9912 and 0.9825) without any salient clustering space. It means that we can not group the prosodic boundaries into some definite levels as we have imagined before. Theoretically you can classify as many levels as you want on a linear line.

However, we let a phonetician to annotate the prosodic boundaries of all these 24 utterance in 4 levels as in C-ToBI, and try to cluster the PBS of the delexical set into 4 clusters too, as shown in table 4 and 5. We surprisingly found that the clustered result agrees very well with the professional annotator except utterance 9. So we can draw some conclusions from the clustering as following: (1) The prosodic boundaries annotated by experts are highly correlated to those done by the native listeners through clustering the results of delexical into the same levels. So a learning sample set for prosodic break annotation was produced from the clustered result as shown in table 5. (2) The perceived breaks form a continuum, which can produce numerous boundary levels theoretically and a blurry part can exist between each two adjacent levels. So a mark can be used to identify the uncertain PBS in C-ToBI.

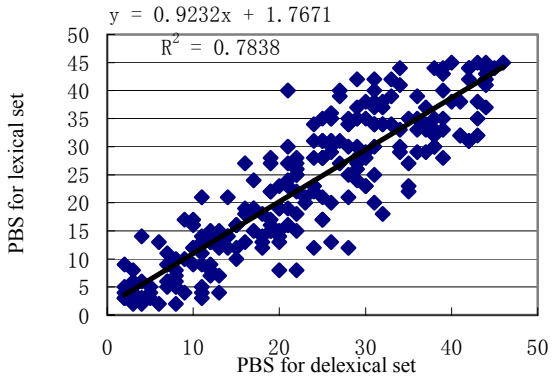

Fig 3: correlation between lexical and delexical sets for all listeners

	Cluster			
	1	2	3	4
AVE	36.78	27.43	15.06	6.27

Table 4: Final Cluster Centers

	By phonetician	Clustered PBS into 4 groups
Syllable boundary within PW BR=0	15, 16, 17	15, 16, 17
PW boundary BR=1	5, 7, 8, 10, 11, 13, 18, 22, 24	5, 7, 8, 10, 11, 13, 18, 22, 24
MIP boundary BR=2	1-4, **9,** 12, 19	1-4, 12, 19
MAP boundary BR=3	6, 14, 20, 21, 23	6, **9,** 14, 20, 21, 23

Table 5: Utterances annotated by phonetician (middle column) and clustered by PBS for de-lexical set (right column)

2.4 Prosodic boundary annotation and acoustic cues analysis

In spontaneous speech, the prosodic boundaries cover more complicated situations than what we have selected in the 24 utterances. The break can be caused by disfluency,

ISBN 1-876346-48-5 © 2003 UAB

filled pause or repetition,etc. All these "abnormal" breaks are annotated with P attached to the break index if the speaker uses a continuous F0 tune even if a strong PBS exists. Meanwhile, the phenomena are labeled on miscellaneous tier with time alignment. Paralinguistic or nonlinguistic phenomena, such as cough, breath, sucking mouth, and background noise, are also annotated in miscellaneous tier. By now a 4 hours' spontaneous corpus has been manually annotated with time aligned segmental and prosodic information by using C-ToBI. The statistic analysis has been made on prosodic boundaries and stress which is partly reported in another paper in this proceeding.

3. CONCLUSION

The answers we get from the perception experiment are
(1) The untrained listeners agree very well with each other in perceptual judgments in lexical set as well as in delexical set, revealing that they can give reliable PBS in annotation task.
(2) It is highly correlated for each subject between two sets, which shows that there is no difference in prosodic boundary annotations between intelligible and unintelligible phrases. In other words, the listeners may judge prosodic boundaries with no reference to semantic, syntactic and lexical information, which proves that current method of prosodic annotation is feasible and scientific.
(3) The prosodic boundaries annotated by experts are highly correlated to those done by the native listeners through clustering the PBS into the same levels. So a learning sample set for prosodic break annotation was produced from the clustered result.
(4) The PBS forms a continuum, which can produce numerous boundary levels theoretically and a blurry part can exist between each two adjacent levels. So uncertainty can exist for boundary annotation.
Results 2 and 3 are similar to what have been found in Dutch [8], so the boundary perception for tonal language is the same for toneless language. But the prosodic cues associated with each boundary lever are not equivalent [5, 19].

ACKNOWLEDGMENTS

The author appreciates Zhigang Yin and Tianqing Wang for their great help in the experiment.

REFERENCES

[1] Aijun Li, Zhigang Yin, et.al., "Spontaneous speech corpus CADCC and it phonetic research", in the *proceedings of the 5th NCMP*, 2001.
[2] Aijun Li, "Chinese prosody and prosodic labeling of spontaneous speech," in *proceedings of speech prosody 2002*.
[3] Beáta Megyesi, Sofia Gustafson-Čapková, "Production and perception of pauses and their Linguistics context in read and spontateous speech in Swedish," in *proceedings of ICSLP'2002*, 2153-2156.
[4] Eleonora Blaauw, "The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech," *Speech Communication* 14 (1994) 359-375.
[5] Maocan Lin, "Sentence break and the prosodic structure", in *Contemporary Linguistics*, Vol. 2, 2000.
[6] Maolin Wang, "A study of pitch of spontaneous speech," *Report of Phonetic Research, Phonetics laboratory*, CASS. 2002.
[7] Hongjun Wang, "Chinese prosodic word and prosodic phrase," *Zhongguo Yuwen*, Vol. 6, 2000.
[8] Jan Roelof de Pijper and Angelien A. Sanderman, "On the perceptual strength of the prosodic boundaries and its relation to super segmental cues," *JASA*, 96(4),Oct. 1994.
[9] Julia Hirschberg and Pilar Prieto, "Traning intoamtional phrasing rule automatically for English and Spanish text-to-speech," *Speech Communication* 18(1996) 281-290.
[10] Lehiste, I., "Perception of sentence and paragraph boundaries," in *Frontiers of speech communication research*, edited by B.Lindblom and S. Öhman (Academic New York), pp. 191-201.
[11] Lehiste, I., and Wang,W.S-Y, "Perception of sentence and paragraph boundaries with and without semantic information," in *Phonologica* 1976, edited by Dresslen and Pfeiffer (Institut für Sprachwissenschaft der universität Innsbruck), pp. 277-283,1977.
[12] Mary E. Beckman & Gayle M.Ayers, "Guidelines for ToBI Labeling," *Manuscript*, Ohio State University, 1994.
[13] Petra Hansson, "Perceived boundary strength," in *proceedings of ICSLP'2002*, 2277-2280.
[14] Pitrelli, John; Beckman, Mary; & Hirschberg, Julia, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," *Proceedings of the 1994 International Conference on Spoken Language Processing*, vol. 1, pp. 123-126. 1994.
[15] Silverman, Kim; Beckman, Mary; Pitrelli, John; Ostendorf, Mari; Wightman, Colin; Price, Patti; Pierrehumbert, Janet; & Hirschberg, Julia, "ToBI: a standard for labeling English prosody," *Proceedings of the 1992 International Conference on Spoken Language Processing*, vol. 2, pp. 867-870.
[16] Selkirk, Elisabeth, " Phonology and syntax: the relation between sound and structure." *Cambridge*, Mass.: MIT Press.1984.
[17] Shih, Chilin, "Mandarin third tone sandhi and prosodic structure." In Wang Jialing and Norval Smith (eds.), *Studies in Chinese Phonology*, 81-123. Berlin: Mouton de Gruyter.1997.
[18] Wightman, S., Shattuck-Hufnagel, S., Ostendorf, M., and Prince, P., "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Am*. 91, 1701-1717, 1992.
[19] Yabin Liu and Aijun Li, "A contrastive study for read and spontaneous speech," *Zhongwen Xinxi Xuebo*, Vol. 1, 2002.
[20] Ziyu Xiong, "A study on the prosodic features of the utterance boundary in natural speech and the interactive function of these features", *Report of Phonetic Research, Phonetics laboratory*, CASS. 2002.