A Perceptual Study on Variability in Break Allocation within Chinese Sentences

Min Chu [†] *Honghui Dong*[‡] *Jianhua Tao*[‡]

[†] Microsoft Research Asia, Beijing

[‡] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing

[†] minchu@microsoft.com; [‡] {hhdong, jhtao}@nlpr.ia.ac.cn

Abstract

This paper investigates the variability of break allocations within Chinese sentences by perceptual experimentation. The results confirm the existence of prosodic chunks. We have found that (1) prosodic chunks are the basic units in the rhythmic organization of Chinese utterances (breaks can generally be allocated by chunk boundaries and breaks placed within a chunk will significantly decrease the naturalness of synthesized speeches); (2) given prosodic chunks, multiple break solutions are acceptable. Furthermore, breaks can be allocated by chunk boundaries using simple rules that impose a length-balance constraint without considering the syntax or semantic structure of a sentence.

1. Introduction

Breaking sentences into suitable rhythmic units is important to achieve high naturalness in speech synthesis. Various machine learning algorithms have been employed to predict the most likely positions for breaks in a text stream [1-4]. These works share an assumption that a unique best-breaking solution exists for any sentence. However, this is not true in many situations. In a recent study, Chu, *et al* [5] reported the variability in the rhythmic organization of a sentence. This implies that there may exist more than one acceptable ways to break a sentence. They found that the variability of higher level prosodic units is larger than that of lower level ones. They also reported the existence of a relatively stable prosodic unit that normally cannot contain an internal break. Such stable units can be grouped into prosodic phrases in various ways. Since the functionality of such stable rhythmic units in prosody is rather similar to that of chunk in natural language processing [6], we refer to it as a *prosodic chunk* in this paper.

After adopting the concept of prosodic chunks, the rhythmic organization of sentences are then decomposed into two steps as shown in Fig. 1. In step one, words are grouped into prosodic chunks by identifying the no-break positions. According to Chu, *et al*. [7], local syntactic structure and length constraints play important roles in prosodic chunking. In step two, prosodic chunks are grouped into prosodic phrases with perceptible breaks in between. We believe that: (1) the operation in step one should be precise, especially on the recall of no-break positions, because any break inserted to a no-break location will significantly decrease the naturalness of synthesized speeches; (2) the operation in step two may be approximate because there are many acceptable ways to group prosodic chunks into phrases. We even believe that the grouping of prosodic chunks can be performed without considering a sentence's syntax structure. Instead, the length-balance constraints are more critical, i.e. prosody phrases in an utterance tend to be of similar lengths. To verify these ideas, we designed a perceptual experiment to evaluate the naturalness of speech synthesized from the same text but different in breaking policies. Two types of utterances were generated: one had breaks at prosodic chunk boundaries (breaks were allocated with simple rules imposing the length-balance constraints without considering sentences' syntactic structure); the other had breaks within prosodic chunks.

The layout of this paper is as the follows: The design and the implementation of the perceptual experiment are introduced in Section 2. The results and analyses are presented in Section 3. Section 4 details our conclusions.

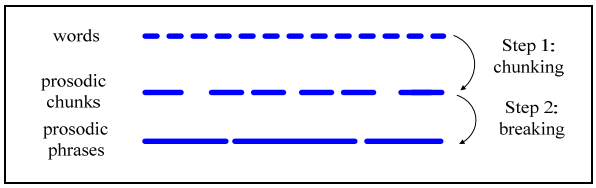

Figure 1: *Rhythmic organization of an utterance*

2. The perceptual experiment

Two hypotheses are tested in the experiment: One is that breaks are acceptable only when they are inserted at prosodic chunk boundaries; the other is that, given prosodic chunks, many acceptable ways to break an utterance exist and breaks can be allocated with simple rules that impose length-balance constraints without considering a sentence's syntactic or semantic structure.

2.1. Design of the experiment

Twenty sentences with 13~19 syllables were selected from the same speech corpus used in [5]. Prosodic chunks had been labeled in the previous study. These sentences contain 4-6 chunks each. Break indices have been labeled to reflect the prosody realization in the recorded speech from a single speaker. Such break solutions for the twenty sentences are referred to as the *natural breaks* in this study. Seven additional break solutions were generated with rules listed in Table 1, five with breaks at chunk boundaries and two with breaks within a chunk. Since the natural break and the five solutions at chunk boundaries support our hypotheses, they are referred to as *positive sample*s in this paper. The other two are then referred to as *negative samples*. As a result, eight break solutions were generated for each sentence, including six positive samples and two negative ones. Details of the eight samples are described in Table 1 with examples. All the eight samples of each sentence were fed into Mulan TTS system [8] with fixed break positions in the input text. All together, 160 utterances were synthesized.

There are at least two ways to compare the naturalness among the eight samples of each sentence. One is to ask

subjects to give a direct MOS score to each utterance. The other is to rank these utterances by one-to-one comparison, by doing AB test. Since the only difference among the eight samples of each sentence is the position of the breaks, obtaining MOS scores precise enough to distinguish their naturalness is not easy. Therefore, we choose the second method, i.e. compared the eight utterances to one another and asked the subjects to decide which utterance sounded better after listening to a pair of utterances. 28 pairs of utterances were generated for each sentence and altogether 560 pairs were obtained for the 20 sentences.

2.2. Experiment procedure

The 560 utterance pairs were randomly sorted and separated into two parts. Subjects were asked to finish the two parts with a not-less-than 30-minute break in between. The utterance pairs were played to the subjects with a scoring tool in a standard PC and subjects listened to them through headphones. The sequence of stimuli played to each subject was randomly generated. Subjects were allowed to listen to each pair as many times as they wanted before they chose either "A sounds better" or "B sounds better". It was suggested that they hear the pairs no more than three times. All together, 20 university students with normal hearing participated in the experiment.

2.3. The preference rate of an utterance

In the experiment, each utterance is compared with seven other samples of the same sentence and was rated by the 20 subjects. If it was judged as the better one in one comparison by one subject, it received one point. Otherwise, it got no points. The ratio between the points an utterance obtained and the number of times it was compared is defined as the preference rate (PR) of the utterance. The PR of an utterance can be calculated on different scales as described in the following sections.

Table 1: *Rules and examples for positive and negative samples*[1].

Positive samples	Rules to generate the samples	Examples
P1	No break in the whole sentence	这个惊人的数字凝聚着一种民族精神。
P2	Insert only one break at the chunk boundary closest to the middle point of the sentence on the left	这个惊人的数字 | 凝聚着一种民族精神。
P3	Insert only one break at the chunk boundary closest to the middle point of the sentence on the right	这个惊人的数字凝聚着 | 一种民族精神。
P4	Insert two breaks at chunk boundaries that are closest to the 1/3 and 2/3 points of a sentence on the left [2]	这个 | 惊人的数字凝聚着 | 一种民族精神。
P5	Insert two breaks at chunk boundaries that are closest to the 1/3 and 2/3 points of a sentence on the right	这个惊人的数字 | 凝聚着一种 | 民族精神。
P6	Use the natural breaks	这个 | 惊人的数字 | 凝聚着一种 | 民族精神。
Negative samples		
N1	Insert one inner-prosodic-chunk break at the left part of the sentence [3]	这个惊人的 | 数字凝聚着一种民族精神。
N2	Insert one inner-prosodic-chunk break at the right part of the sentence	这个惊人的数字凝聚着一种民族 | 精神。

3. Results and Analyses

3.1. Consistency among sentences

Before a formal analysis, the consistency of PRs among the 20 sentences was examined. To do this, the PR was first calculated for each utterance within seven pairs from the scores given the 20 subjects. As a result, an eight-by-eight matrix was obtained for each sentence. Table 2 provides an example. The number in each cell shows the preference rate of the utterance on the left to the utterance on the top. Since each utterance hasn't been compared with itself in the experiment, the diagonal cells are empty. Then, the average of the 20 matrixes is obtained and the correlation coefficients between individual matrixes and the average matrix are calculated. The results are shown in Fig. 2. The correlation coefficient of a sentence reveals its consistency with other sentences.

Table 2: *An example of the PR matrix of one sentence calculated from the scores given by the 20 subjects.*

	P1	P2	P3	P4	P5	P6	N1	N2
P1	-	0.4	0.8	0.65	0.6	0.75	0.9	1
P2	0.6	-	0.6	0.75	0.65	0.95	0.85	0.85
P3	0.2	0.4	-	0.5	0.45	0.65	0.95	0.9
P4	0.35	0.25	0.5	-	0.45	0.75	0.7	0.8
P5	0.4	0.35	0.55	0.55	-	0.65	0.8	0.85
P6	0.25	0.05	0.35	0.25	0.35	-	0.75	0.45
N1	0.1	0.15	0.05	0.3	0.2	0.25	-	0.55
N2	0	0.15	0.1	0.2	0.15	0.55	0.45	-

From Fig. 2, it is seen that sentence 5, 6 and 20 have much lower correlation coefficients than the others. When looking into their PR matrixes, we find that both 5 and 20 contain a positive sample that should be treated as a negative one because they all have a break shifted one syllable from the desired position by mistyping in the data preparation. So,

[1] The prosodic chunk boundaries in the example sentence are marked with "#" as "这个#惊人的数字#凝聚着#一种#民族精神。"and the locations of breaks are marked with "|" in Table 1

[2] In P4 and P5, if no boundary on the left or right is found, the one on the right or left can be used. We have made sure that P4 and P5 differ at least in one break position.

[3] In N1 and N2, breaks are still put at prosodic word boundaries, but they are within a prosodic chunk.

they are moved from the pool of positive samples to that of negative samples in the following analyses. The results for sentence 6 look strange because several positive samples are rated lower than one of the negative samples. When we listen to these samples, we cannot agree with the results. Therefore, we asked four new subjects with normal hearing to score the 28 pairs of this sentence again. Their results confirm our intuitive analysis. Thus, we decided to remove sentence 6 in the remaining analyses.

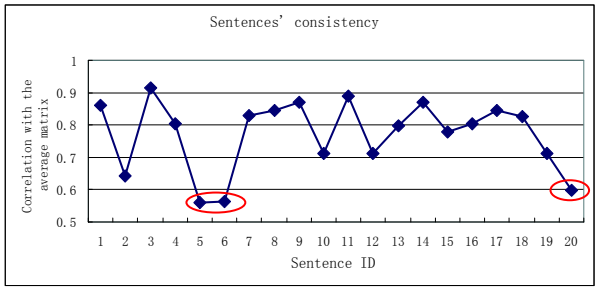

Figure 2: *Consistency among 20 sentences*.

3.2. Consistency among subjects

In order to achieve reliable results, consistency among subjects was investigated as well. The PR matrix for a subject was calculated from his/her scores for the 20 sentences. The correlation coefficients between individual matrixes of all subjects and their average matrix, as shown in Fig. 3, reveal the consistency of each subject to the others. It is seen that most subjects have similar preference tendencies regarding the eight utterances of each sentence except for subject 4 and 18. The possible reasons may be that they did not have a good understanding of the task or they didn't concentrate on it. We decided to remove their results from our analyses.

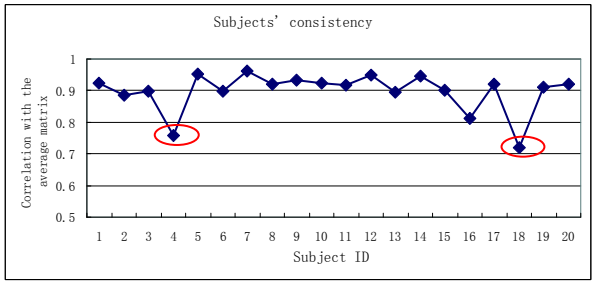

Figure 3: *Consistency among 20 subjects*

3.3. Positive vs. negative samples

After unsuitable utterances and non-qualified subjects were removed, the PRs of each utterance to seven other utterances were recalculated and the average of the seven PRs (denoted as APR) represents the overall preference rate of that utterance. A larger APR implies the corresponding utterance sound better than the others and vice versa. All utterances were classified into two categories, positive and negative samples. The means and standard deviations of APRs in the two categories are shown in Fig. 4. We can see that, generally, positive samples sound significantly better than negative samples (P≈0.00). This result supports our hypothesis about the existence of prosodic chunks, i.e. prosodic chunk boundaries are much better locations to break than non chunk-boundaries.

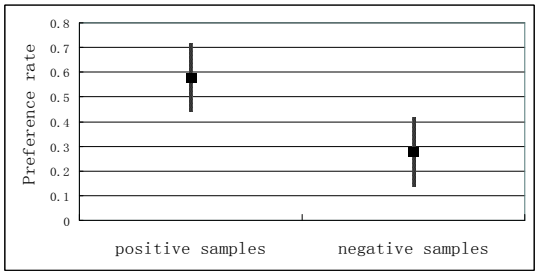

Figure 4: *Means and standard deviations of APRs in positive and negative samples*.

To verify the second hypothesis, we took a close look into the positive samples.

3.4. Preference rate among positive samples

In this section, APRs were recalculated only among positive utterances P1-6. The mean value of the new APRs of all positive samples is 0.50 with a standard deviation of 0.16. This shows that these positive samples sound similar in terms of their naturalness, i.e. the six positive ways to break a sentence are all acceptable.

These positive samples were then classified into four categories according to their breaking rules: 1) no break: contains all P1 utterances (19 samples); 2) single break: contains all P2 and P3 utterances (37 samples); 3) two break: contains all P4 and P5 utterances (37 samples); and 4) natural break: contains all P6 utterances (19 samples). The means and standard deviations of APRs in the four categories are shown in Fig. 5. We can see that the single-break and two-break groups have slightly higher APRs than the other two. However, the analysis of variance (ANOVA) shows that the differences among groups are not significant (P=0.45). It is interesting that, breaks in the single- and two-break categories are inserted with length balance rules without considering syntactic structure and they sound equally good as samples with natural breaks and no break. This observation supports our second hypothesis that, given prosodic chunks, multiple break solutions are acceptable, and syntactic or semantic structure of a sentence does not affect much on the grouping of prosodic chunks into prosody phrases.

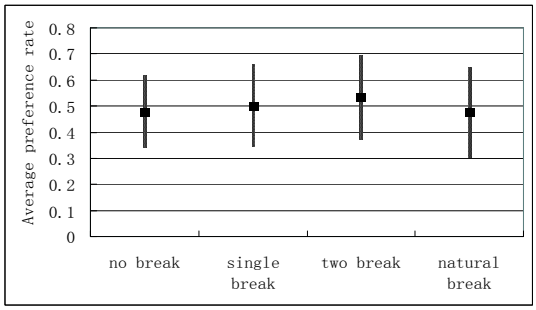

Figure 5: *Means and standard deviations of APRs in four types of positive samples*.

3.5. The length constraints in breaking sentences

Although most positive samples sound similar in terms of their naturalness, some are worse than others. Nine samples with APRs lower than 0.3 were found. One of them is in the no-break type, four in the single-break type and four in the natural-break type. None is in the two-break type. It seems that single-break and natural-break types have greater

problems than others. Therefore, we performed detail analyses of these two types.

The four worst samples in the single-break type are listed in Table 3. We find that one possible reason for the low preference rate is that the length of the two prosodic phrases separated by the break is not balanced (because we couldn't find better location in the single break situation beside the other one). Therefore, a length balance ratio (LBR) is defined for single-break utterances as the ratio between the length difference of the two prosodic phrases in the utterance and the average length of the two phrases, as shown in equation (1).

$$LBR = abs(len1 - len2) * 2 / (len1 + len2) \qquad (1)$$

The relationship between LBR and APR of utterances is shown in Fig. 6. The solid line is the trend line. The correlation coefficient between the two parameters is -0.44. That is to say the APR of an utterance is negatively correlated to its LBR.

Table 3: *Worst samples in the single-break category.*

1	人生最大的痛苦莫过于 | 失去亲人。
2	世界冠军和大学生 | 谈为国奉献。
3	要根据 | 农业的需要发展农用工业。
4	合理利用 |长江岸线是建港的基本原则

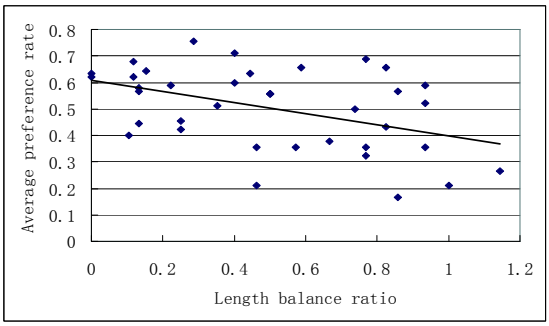

Figure 6: *Relationships between LBR and APR of utterances*

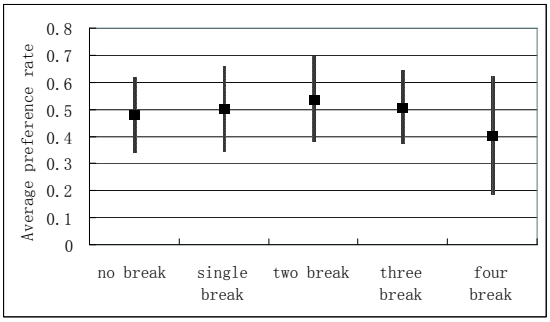

Figure 7: *Means and standard deviations of APRs in groups of utterances with various numbers of breaks.*

Three of the four worst natural-break samples have breaks at all prosodic chunk boundaries. This may imply that two many breaks may hurt the naturalness of synthesized speech although the speaker can do well with the same breaks. Therefore, the natural samples are further classified into two groups in accordance with the number of breaks. In the natural samples, only one utterance has two breaks and it was merged into the two-break category. Eleven utterances have three breaks and six have four. They formed the two new categories. Then, Fig. 5 was converted to Fig. 7, showing the statistics of APRs in groups of utterances with different numbers of breaks. We can see that utterances with two breaks have the highest mean APR and those with four breaks have lower APR than the others. However, in an ANOVA analysis, these differences are not statistically significant. One possible reason is because we have two few samples in the three-break and four-break categories.

4. Conclusions

This paper investigates the variability in allocating breaks within sentences through a perceptual experiment. The results show that prosodic chunks are relatively stable units in the rhythmic organization in Chinese. Breaks allocated to locations within prosodic chunks will significantly hurt the naturalness of synthesized speech. However, variability exists in grouping prosodic chunks into prosody phrases. For most sentences, multiple break solutions are acceptable. Furthermore, length-balance constraint plays a much more important role in allocating breaks to chunk boundaries than the syntactic or semantic structure of a sentence. For a sentence with 13-19 characters, inserting one to two breaks is little more suitable than no break or additional breaks. However, the differences among these solutions were not statistically significant in our results.

Based on above conclusions, we can see that predicting no-break locations in a sentence is much more important than predicting the break locations. After obtaining the prosodic chunks in a sentence, many flexible ways can be used to allocating breaks by chunk boundaries. Our future work will focus on prosodic chunking.

5. References

[1] Wang, M.Q. and Hirschberg, J., 1991. Predicting intonational phrasing from text. *Association for Computational Linguistics 29th Annual Meeting*, 285-292.

[2] Ostendorf, M. and Veilleux, N., 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics* 20(1), pp. 27-54.

[3] Taylor, P. and Black, A.W., 1998. Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language* 12, 99-117.

[4] Chu, M. and Qian, Y., 2001. Locating boundaries for prosodic constituents in unrestricted Mandarin texts. *Computational Linguistics and Chinese Language Processing* 6(1), 61-82.

[5] Chu, M., Zhao, Y. and Chang, E., 2005. Modeling stylized invariance and local variability of prosody in text-to-speech synthesis. accepted by *Speech Communication*.

[6] Abney, S. P., 1991. Parsing by chunks. *Principle-Based Parsing: Computation and Psycholinguistics*, Berwick, R.C., Abney, S.P., Tenny, C.,eds. Kluwer, Dordrecht, 257-278.

[7] Chu, M., Wang, Y.J. and Bao, M.Z., 2004. Local syntactic constraints and length constraints in the rhythmic organization of Chinese Putonghua. *Collection of Linguistic Studies*, Vol. 30, 129-146. (in Chinese)

[8] Chu, M, Peng, H., Zhao, Y., Niu, Z and Chang, E., 2003. Microsoft Mulan — a bilingual TTS systems, *Proc. ICASSP2003*, Hong Kong.