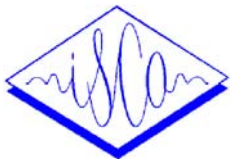

Speech Prosody 2002
Aix-en-Provence, France
April 11-13, 2002

Prosodic Word: the Lowest Constituent in the Mandarin Prosody Processing

Yao Qian & Wuyun Pan

Linguistics Institute
Shanghai Normal University, China
{yqian; pwy}@shtu.edu.cn

Abstract

This paper proposed a novel method, which is using prosodic word as the lowest constituent in the prosody processing, to solve the prosody problem of Mandarin concatenative speech synthesizer based on a large corpus. The results, obtained from applying new solution to deal with the intonational prominence placement and break boundaries assigning in text-to-speech systems, are positive and encouraging.

1. Introduction

Data-driven Mandarin text-to-speech systems can be able to produce more natural synthesized speech than the others. They base on an ultimate assumption that they have a very large speech corpus containing enough prosodic and spectral varieties for all synthetic units [1]. When synthesize new text, the system will select the synthetic unit which has the same context as the text inputted, from a very large corpus. So the Data-driven speech synthesizers meet the difficulties of recording data covering and synthetic units selecting. The phonetic and prosodic knowledge is very useful for them. Syllable is the smallest unit normally used for Mandarin speech concatenation. The varieties of syllable spectra and prosody have large relationship with its prosodic and phonetic context information, such as the left and right syllables, the position in word and phrase.

Chinese texts do not contain any visual clue boundaries. Word segmentation becomes a basic requirement for almost all text analyses. Many studies had been done on word segmentation. Besides, Chinese has no distinct boundary between phrase and word. Some characters combination is not only a word, but also a minor phrase. However, in spoken Chinese, there exists a disyllable rhythm (or prosody). In order to meet natural and beautiful prosody, Succeeding mono-character words are often uttered as one disyllabic unit of rhythm and long words (may be a phrase in other sentence) are often uttered as several units. For example, in a Chinese sentence, “我买了一本好书 (I brought a good book)”, each character itself is a lexical word (L-word). Yet, in natural speech, the basic units of rhythm are “我”, “买了”, “一本” and “好书”. The unit of rhythm in Mandarin is referred as prosodic word(P-word), which is defined as a group of syllables that should be uttered closely and continuously.

According to the human perception, the understanding of big language unit is based on that of small unit. So there are many benefits from the hierarchical solution, such as flexibility and controllability. In this paper, we put forward a solution using prosodic word as the lowest constituent in the Mandarin prosody processing. When employed to solve the issues of phrase accent placement and prosodic constituent locating, it achieved high performance in perceptual experiment.

Section 2 illustrates the importance of the segmenting unrestricted Chinese text into P-words instead of L-words and how to do it. Section 3 shows the P-word application in solving the prosody issues for Mandarin text-to-speech system. Section 4 gives the conclusion and discussion.

2. Prosodic word

2.1. Annotating prosodic word boundaries

P-word is the basic prosodic unit in Mandarin speech. It is formed dynamically according to the context. Many possible combinations of characters exist in different real texts. So it’s impossible to list all P-word in a lexicon as what is done for L-word. In order to find the form rules of the P-word, we’ve annotated P-word boundaries in some corpus by listening to the utterances and reading the text transcriptions. In other words, labeling P-word boundaries based on perception aided by general linguistics knowledge that are mainly listed in following:

- FUNCTION, or CLOSED words, such as prepositions and articles, are looked as clitic. A disyllabic or tri-syllabic L-word is a P-word if it has no proclitic or enclitic. Otherwise, it forms a P-word with its clitic. Examples for enclitic are “的、了、着、（楼）上、（地）下、（物理）学、（革命）性”; Examples for proclitic are “副（所长）、半（正式）”.
- A mono-syllabic L-word often forms a P-word with the L-word before or follow it. Only when a mono-syllabic L-word is lengthened long enough to balance the disyllabic rhythm, it becomes a mono-syllabic P-word.
- All L-word contain more than 3 syllables should be segmented into several disyllabic or tri-syllabic P-word according to their structures. When there have proclitics or enclitics, the clitics merge into the first or last P-word in the long L-word.

A large speech corpus, which contains 11248 sentences, has been collected and annotated. The length of these sentences is between 10 and 30 characters. P-word boundaries are annotated manually in the script of the corpus. In exploratory experiment, 1348 sentences are annotated three times by three annotators (HJY, ZF and ZR) separately and the resulted three annotations are compared in table 1, where precision and recall are given by

$$precision = CPWB / APWB * 100\% \quad (1)$$

$$recall = CPWB / ARPWB * 100\% \quad (2)$$

10.21437/SpeechProsody.2002-133

where, ARPWB, standing for all real P-word boundary, is the total number of real P-word boundaries. (If more than two annotators share the same opinion on the location of a boundary, the boundary is kept as a real one). APWB, standing for annotated P-word boundary, is the total number of P-word boundaries annotated by an annotator and CPWB (correct P-word boundary) is the number of boundaries annotated correctly by the annotator. From table 1, we find that very high ratio of agreement on the locations of P-word boundaries has been achieved among the three annotators. The remaining sentences are annotated only once by them to reduce workloads. A total of 77642 P-words are annotated. They are used as the ARPWB reference.

*Tabl*e 1. Precision and recall on P-word boundaries for three annotators.

Annotators	HJY	ZF	ZR
Precision (%)	98.9	98.5	99.3
Recall (%)	99.2	99.3	98.9

2.2. Prosodic word vs. lexical word

All sentences in the script for the speech corpus are segmented into L-words by a block-based robust dependency parser [2]. Totally 95831 L-words are obtained. This number is 23.4% larger than that of P-word. A P-word can contain more than one L-word and it can also be only a part of a L-word. If all the L-words are judged as P-words, we get 70.71% and 93.62% for the precision and recall respectively, which reveal the great differences between P-word and L-word. The distribution of length of P-words and L-words in the corpus is shown in Figure 1, from where we find that there are much more mono-character L-words than P-words and more bi-character P-words than L-words. The maximum length of P-word in the corpus is 5-character, while, the maximum length of L-word in the corpus is 13-character.

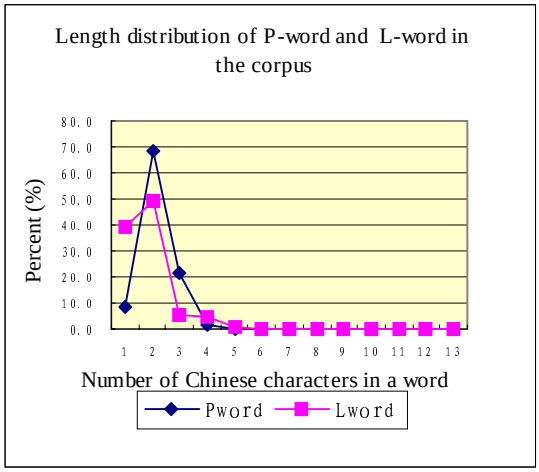

Figure 1. Length distribution of P-word and L-word in the corpus.

2.3. Predicting p-word boundaries

A very important feature for P-word that discriminates it from L-word is that it is constrained not only by semantic requirement of a sentence, but also by the physical mechanism of articulators and the beauty of rhythm in speech. If all L-words longer than 3 characters are splitted into several shorter P-words, the precision and recall rates increase to 71.69% and 98.8% respectively, which is used as the reference performance for our P-word segmentation methods. The splitting of longer L-word is realized by adding structural information into the lexicon. After performing the splitting, enough high recall is obtained, yet, the precision is far from satisfaction. However, P-word strings can be predicted from L-word string [3]. Many features driven from text, such as Part-of-speech tagging of L-word, the length in characters of L-word and word position in sentence, are employed in training P-word boundary prediction. It achieved 92.41% of precision and 94.46% of recall on testing data respectively.

3. The lowest constituent in the mandarin prosody processing

3.1. The potential break boundary site

P-word is looked as the lowest constituent in the prosodic hierarchy and should have a perceivable prosodic boundary. In real speech, not all boundaries of P-word have breaks, it is tolerable if there is a break at the boundary of the P-word. Any inner P-word break will make the speech unintelligible or unnatural. So every P-word boundary is the potential break boundary site (PBS).

There are many linguistic literatures specifying various hierarchical structures for prosodic constituents. Intonational phrase (INP) and intermediate phrase (IMP) are the most commonly accepted levels in English. An English sentence consists of a sequence of INP and each INP, in turn, is composed of a sequence of IMP. In perception, the INP boundaries are perceived by major break and IMP boundaries are perceived by minor break. According to the analysis of P-word, A three-tier instead of the conventional two-tier prosodic hierarchy is defined for a sentence in Mandarin. We add P-word into Mandarin prosodic hierarchy as a lowest constituent. A sentence consists of one or more INP. An INP is decomposed into several IMP and the building blocks for an IMP are P-word. An INP boundary necessarily coincides with an IMP boundary and an IMP boundary is an inevitable P-word boundary, but, not vice versa. The acoustic cues to INP and IMP boundary are major silence and minor silence. In addition, The duration of final syllable of the phrase is lengthened by speaker [4].

When automatically locating boundaries for prosodic constituents in unrestricted Chinese text, a bottom-up hierarchical approach is proposed [5]. IMP boundaries are detected only from PBSs that are judged as P-word boundaries. Then, INP boundaries are picked up only from the predicted IMP boundaries. This hierarchical processing method is more effective than that of predicting the all boundaries at one time. Compared with the result manually annotated, the result automatically annotated has 82.49% of overall accuracy on testing data.

There is randomicity for breaking when people speak. A perceptual experiment is used for the performance evaluation from the perceptual point of view. Speech waves are synthesized with Microsoft data-driven TTS system, which takes in two types of inputs:

- Type A: sentences with P-word boundaries generated automatically.

- Type B: sentences with L-word boundary only.

Two-version speech waves of total 108 sentences picked up from the testing set are synthesized. And 2 comparing pairs (AB and BA) are formed for each sentence. Totally 15 subjects take part in the experiments, each of them listens to part of these comparing pairs and is forced to select a better utterance in each pair. The preference rate is counted as:

$$P_T = count(T) / \sum count(T) \quad \text{, T=A or B} \tag{3}$$

where, P_T is the total number of times when type T is selected.

The final preference rates for all two types are shown in figure 2. It can be found that type A (the automatically generated P-word string) sounds much better than type B (L-word strings). This result elucidates the importance of regarding P-word as a lowest constituent in prosodic hierarchy.

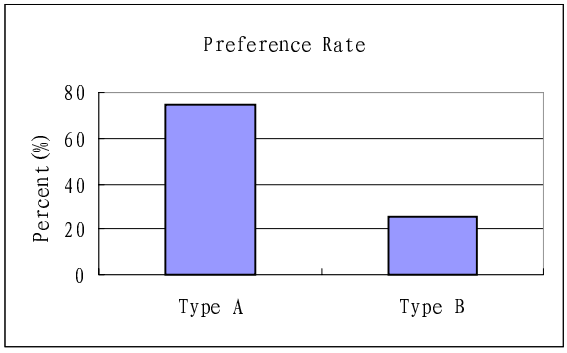

Figure 2: Preference rates for two types of synthesized speech. Type A, synthesized from automatically annotated P-word strings; Type B, synthesized from L-word strings

3.2. The potential unit bearing intonational prominence

In Chinese, Speakers make some words more prominent in intonation than the others within an international phrase. These words are said to be accented or to bear phrase accent. Sentence accent is most prominent word in the sentence. In general, the sentence accent often occurs on the phrase-accented word. Phrase accent placement becomes important to improve TTS naturalness and intelligence after locating prosodic constituents successfully.

Two single speaker read speech corpuses, One speech corpus (sentence corpus) is made up of 3000 sentences, The other (discourse corpus) contains 21 discourses about 2000 sentences and 67 minutes speech in total, are assigned the phrase accent by annotators who have linguistics background through listening to the corpus. After analying the corpus annotated, we find that only 30.9% and 35.8% of the words in two speech corpuses are accented. The phrase accent more likely occurs to P-word. Moreover, the boundaries of accented syllable group always coincide with those of P-word. Since the information of P-word boundary is useful to predict the phrase accent, P-word is looked as the potential unit bearing intonational prominence.

How human decide which words to accent and which to deaccent—what constrains accent placement and what function accent serves in conveying meaning—is an open issue in linguistic and speech science. In general, syntactic structure, semantic, and discourse/pragmatic factor are believed to determine accent placement. However, these analyses for unrestricted text in TTS system cannot be processed automatically in real time, while need high accuracy. There only have many domain-specific systems that are capable of meeting these requirements up to now. Currently, there have been new and successful efforts to find ways of using word class, surface position, FOCUS and the GIVEN/NEW distinction on modeling local text for accent prediction [6] and [7]. So we use Part-of-speech, prosodic boundaries, word position, word unigram score and TF-IDF weight (be widely used to qualify the word importance in information retrieval tasks) in two speech corpus training by machine learning method. It achieved the 80.01% and 77.15% of accuracy on testing data respectively [8].

Though the phrase accent annotated manually can be used as the reference for evaluating the results generated automatically, but there are some arbitrariness for phrase accent placement when people speak. So the manually annotation isn't the only criterion for the performance evaluation. We designed a perceptual experiment to evaluate the phrase accent assigned automatically, which is the same as that of breaks assigning. Since the longer duration is the acoustic cue for phrase accent [9], syllable duration are added into our data-drive concatenative speech synthesis system as one of the factors considered for selecting the candidate units. We also enlarge the amplitude of concatenative units that will bear the phrase accent. Speech waves are synthesized with Motorola TTS system [10], which takes in two types of inputs:

- Type A: sentences with phrase accent generated automatically .
- Type B: sentences without phrase accent.

Totally 50 sentences are picked up and 8 subjects take part in the experiment. The final preference rate for all two types is shown in figure 3, where we can find that the preference rate of Type A (automatically annotated) increases 20 percent relatively compared with Type B (no annotation). We think this improvement is remarkable due to small coverage of accented word in Chinese. The result elucidates the importance of assigning phrase accent for Chinese TTS system. It also shows that using P-word as the potential unit bearing phrase accent is an effective way.

4. Discussion and conclusion

This paper discussed the solution of prosody processing in Data-driven Mandarin text-to-speech systems. The method of using prosodic word as the lowest constituent proved to be highly effective. Prosody belongs to perception category. Automatic prosody prediction needs a large corpus with annotation, which will be used as training data. How to assure quality equalization among different annotator and whether automatic labeling prosody according to acoustic cues, will need further research.

5. Acknowledgments

This paper is supported by Microsoft Research China and Motorola China Research Center. The authors are especially grateful to Min Chu and Fang Chen for providing the Text-

To-Speech systems. The authors thank everybody who takes part into the perceptual test.

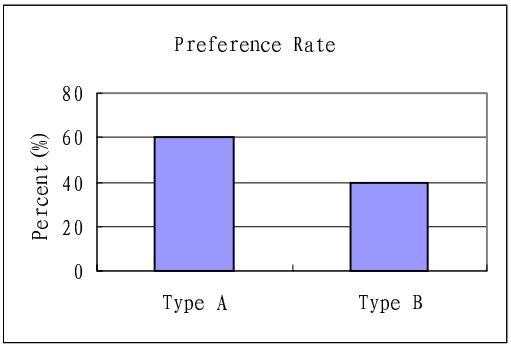

Figure 3. Preference rates for two types of synthesized speech.Type A, synthesized from automatically annotated phrase accent string; Type B, synthesized from string without phrase accent annotation.

6. References

[1] Chu, M.; Peng, H.; Yang, H.; Chang, E., 2001. Selection Non-uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer, *In Proceedings of 26th International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City.*

[2] Zhou, M., 2000. A Block-based Robust Dependency Parser for Unrestricted Chinese Text, *In the second Chinese Language Processing Workshop attached to ACL2000, Hong Kong.*

[3] Qian, Y.; Chu, M., 2001. Segmenting Unrestricted Chinese Text into Prosodic Words Instead of Lexical Words, *In Proceedings of 26th International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City.*

[4] Qian , Y.; Chu, M.; Pan, W., 2001. The Acoustic Cues to Mandarin Prosodic Constituents, *In Proceedings of 5th national Conference On Modern Phonetics, Bei Jing.*

[5] Chu, M.; Qian, Y., 2001. Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts, *accepted for publication in International Journal of Computational Linguistic & Chinese Language Processing.*

[6] Hirschberg Julia, 1993. Pitch Accent in Context: Predicting Intonational Prominence from Text , *Artificial Intelligence.*

[7] Pan Shimei; Kathy McKeown, 1999. Word Informativeness and Automatic Pitch Accent Modeling , *In Proceedings of EMNLP/VLC'99, College Park, Maryland.*

[8] Qian, Y.; Chen, F., 2002. Assigning Phrase Accent to Chinese Text-to-speech System, *accepted by International Conference on Acoustics, Speech, and Signal Processing.*

[9] Ma, M., 1998. Weak stress pattern in Mandarin, *M.S. thesis in Linguistics Institute of SHTU.*

[10] Chen, F.; Chen, G.; Huang, J.; Yu, Z.; Yue, D.; Zu, Y., 2001. Natural Sounding Embedded Text-To-Speech Systems, *In Proceedings of 5th National Conference on Modern Phonetics, Bei Jing.*