High-Quality Prosody Generation in Mandarin Text-to-Speech System

● Qing Guo ● Jie Zhang ● Nobuyuki Katae ● Hao Yu

(Manuscript received March 19, 2009)

A text-to-speech (TTS) synthesizer is a computer-based system that can automatically read text aloud. Fujitsu is developing a Mandarin TTS system using state-of-the-art technologies. The prosodic structure of synthesized text provides important information for making synthetic speech produced by a TTS system more natural and understandable. This paper describes a global probability estimation method for predicting prosodic words, which are the lowest constituent of the prosodic structure. Experimental results for this method are very promising. They are better than those for our previous binary prosodic tree method in terms of both accuracy and memory cost.

1. Introduction

A text-to-speech (TTS) synthesizer is a computer-based system that can read text aloud automatically, regardless of whether the text is introduced by a computer input stream or a scanned input submitted to an optical character recognition engine. TTS synthesis can be used in many areas, such as telecommunication services, language education, vocal monitoring, and multimedia, and it can also be used as an aid to handicapped people. Impressive progress has been made over the last couple of decades in Mandarin TTS research. A lot of mature high-quality Mandarin TTS systems are now in commercial use.

Concatenative speech synthesis using a large speech database has become popular in recent years due to its improved sensitivity to unit context over simpler predecessors. These systems usually make use of large speech databases and sophisticated search algorithms to determine the optimal unit sequence for synthesizing each sentence. During synthesis, the input text is converted into a phonetic string with predicted prosodic targets. The synthesis system then searches for speech segments that are close to the context of the target phonetic string and its predicted prosodic targets. Finally, a pitch and duration modification algorithm, such as PSOLA, is applied to pre-stored units to guarantee that the prosodic features of the synthetic speech meet the predicted target values, and the transformed waveforms are then concatenated together.

Rhythm is an important factor that makes the synthesized speech of a TTS system more natural and understandable. Researchers have found that there is a hierarchical prosodic structure for Chinese prosody, which constitutes the rhythm of Chinese speech.[1] The boundaries of prosodic units can be identified by pauses, pitch changes, or duration changes in boundary syllables in the speech. In a TTS system, the prosodic structure provides important information for the prosody generation model to produce all these effects in the synthesized speech.

There are many reports specifying various

hierarchical structures for prosodic constituents. Generally speaking, the main prosodic constituents of Chinese speech are the prosodic word (PW), prosodic phrase, and intonation phrase. A PW is a group of syllables uttered continuously and closely without breaks in the speech. It is the lowest constituent of the prosodic hierarchy and should have a perceivable prosodic boundary. Thus, good PW grouping plays an important role in increasing the naturalness of synthesized speech.

Automatic prosody generators, however, cannot yet deliver high-quality prosody. One of the main obstacles to automatic prosody generation is the difficulty found in identifying the hierarchical prosodic constituents from texts automatically. It has been proven through many experiments that prosody constituents are not always identical to those of the surface syntax. The relationship between prosody and syntax is not well understood.

In this paper, we describe a global probability estimation method based on the positional types of lexical words according to their positions in the PWs that they belong to in order to predict the PW boundary. The rest of this paper is organized as follows. Section 2 outlines the system framework of the Fujitsu Mandarin TTS system. Section 3 briefly introduces the speech database used in our research and then describes the global PW grouping probability estimation method in detail. Section 4 presents experiment results and discusses them. Section 5 concludes with a summary of the main points.

2. Fujitsu Mandarin TTS system

The Fujitsu Mandarin TTS system is a state-of-the-art, unit-selection-based concatenative speech synthesis system. Its framework is shown in **Figure 1**. There are three main modules in the system: the text analysis, prosody generation, and speech synthesis modules. The text analysis module is responsible for text normalization, digital and special symbol tokenization, word segmentation and part-of-speech (POS) tagging, phonological analysis, homograph disambiguation, tone sandhi processing, and prosodic structure prediction. The prosody generation module performs duration prediction and pitch prediction. The speech synthesis module performs unit selection, voice generation, and waveform concatenation.

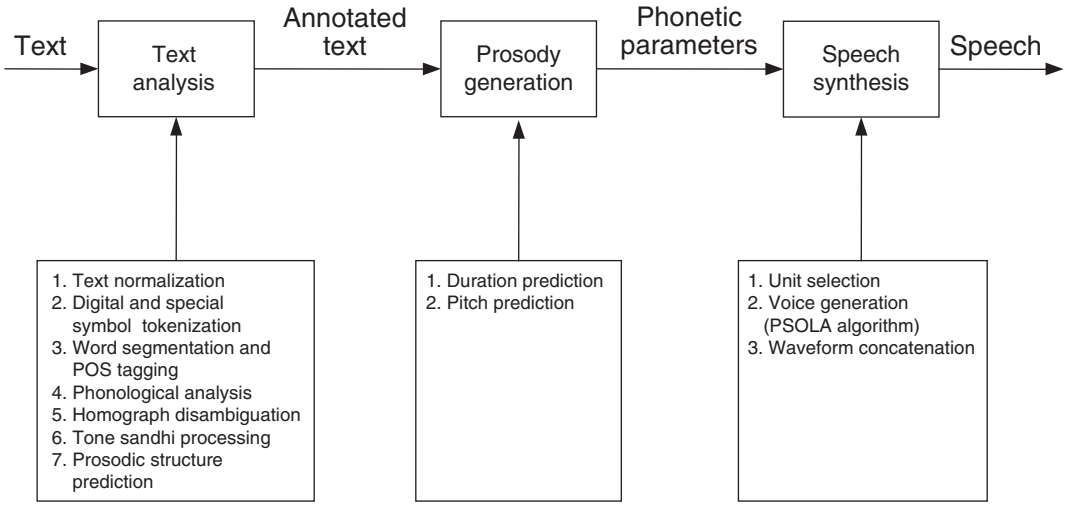

Figure 1
Framework of Fujitsu Mandarin TTS system.

Duration is one of the most important prosodic features contributing to the perceived naturalness of synthetic speech. The variation in segmental duration can hint at the identity of the speech sound and help the segmentation of a continuous flow of sounds into words and phrases, thereby increasing naturalness and intelligibility. The pitch contour is another key parameter that affects the quality of the synthesized speech. Tonal languages, such as Chinese, use variations in pitch to distinguish otherwise identical syllables. Thus, good pitch prediction in a TTS system is important not only for natural-sounding speech but also for good intelligibility.

The Fujitsu Mandarin TTS system uses a decision-tree-based duration modeling method and a statistical pitch contour prediction method.[2] Prosody evaluation results indicate that its prosody generation module generates much better prosody than two other famous Mandarin TTS systems.

As mentioned in the previous section, the performances of both the duration prediction and pitch contour prediction models are highly dependent on automatic identification of the hierarchical prosodic constituents from texts. Therefore, this paper discusses a new PW grouping method based on global probability estimation.

3. Global PW grouping probability estimation

3.1 Speech database

The text in our speech database came from the People's Daily Corpus 1998, which was transcribed from a Chinese newspaper with word segmentation and POS-tagging annotated for natural language processing purposes. 3360 sentences with about 200 000 Chinese characters were selected from the text corpus using a greedy algorithm.

The prosody structure used in this paper is composed of four tiers:[1] prosodic word (PW), minor phrase (MIP), major phrase (MAP), and intonation group (IG). An example of the structure of Chinese prosody is shown in **Figure 2**. PW is a tone group bearing one word stress. MIP contains one or more PWs and bears one phrasal stress, and the perceived break between MIPs is longer than that between PWs. MAP contains one or more PWs and bears one phrasal stress, and the perceived break between MAPs is longer than that between MIPs. The criterion for prosody structure labeling is listening perception. MAPs are often marked by commas with incomplete pitch resetting while IGs are marked by periods, quotation marks, or semicolons with full pitch resetting. In addition, three levels of stress have been defined, namely, stressed, normal, and neutralized.

The following is a sample transcription of a

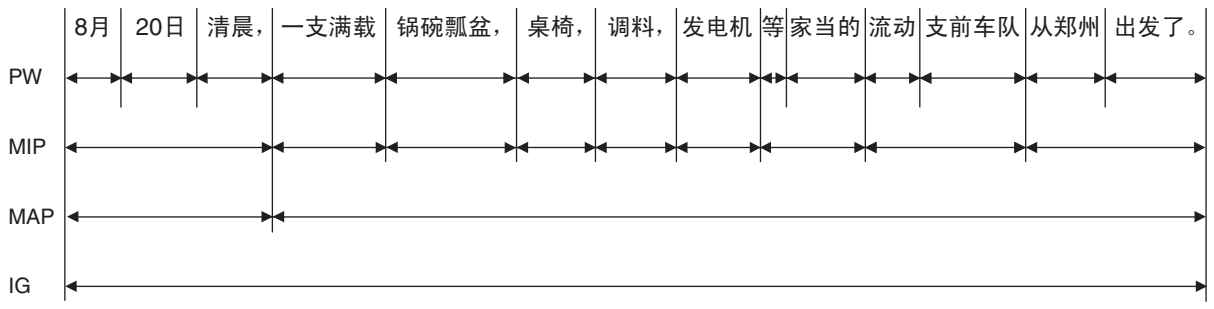

Figure 2
Example of structure of Chinese prosody.

certain sentence in the speech corpus. Here, "|", "||", "|||", and "@" represent PW, MIP, MAP, and IG, respectively, in the transcription. A syllable marked with "_H" means it is a stressed syllable, and a syllable marked with "_L" means it is a neutralized one.

8月(ba1 yve4_H)/t | 20日(er4 sh%2_H r%4_H)/t | 清晨(qing1_H chen2)/t， ||| 一(yi1)/m 支(zh%1_H)/q 满载(man3 zai4_H)/v || 锅碗瓢盆(guo1_H wan3 piao2_H pen2)/l， || 桌椅(zhuo1_H yi3)/n ， || 调料 (tiao2_H liao4)/n ， || 发电机(fa1 dian4 ji1_H)/n || 等(deng3)/u | 家当(jia1 dang4_H)/n的(de5_L)/u || 流动(liu2 dong4_H)/vn | 支前(zh%1_H qian2/vn 车队(che1_H dui4)/n || 从(cong2_H) /p 郑州(zheng4 zhou1_H)/ns | 出发(chu1 fa1_H)/v 了(le5_L)/y 。@

3.2 New PW grouping method

Four kinds of position types are defined for lexical words according to their positions in the PWs that they belong to. These four position types are denoted by B_1, B_2, M, and I, where B_1 means that a lexicon word is at the beginning of the PW that it belongs to, B_2 means that a lexicon word is the second lexical word of the PW that it belongs to, M means that a lexicon word is at the third or another end position of the PW that it belongs to, and I means that a lexicon word belongs to a singleton PW that has only one lexical word. The original transcription of the training sentences can be formatted easily as follows:

The sentence "晚饭/n 后/f ||| 我们/r | 决定/v | 先/d 去/v 逛逛/v || 张家港/ns 的/u | 市容/n 。@" is formatted to "晚饭/n/B1 后/f/B2 ||| 我们/r/I | 决定/v/I | 先/d/B1 去/v/B2 逛逛/v/M || 张家港/ns/B1 的/u/B2 | 市容/n/I 。@"

A sentence with segmented words can be represented by a word sequence as follows:

$W = w_1 w_2 \cdots w_{n-1} w_n$. Let $pos_i, i = 1,2,\cdots,n$ denote the POS of w_i.

One possible PW grouping result PW for the sentence can be written as

$$PW = w_1 s_1 w_2 s_2 \cdots w_{n-1} s_{n-1} w_n,$$

in which $s_i \in \{B_1, B_2, M, I\}$, $i = 1,2,\cdots,n$.

The target of PW grouping is to find the optimum PW grouping PW^* from all possible paths.

$$PW^* = \max_{s_1, s_2, \cdots, s_{n-1}} P(w_1 s_1 w_2 s_2 \cdots w_{n-1} s_{n-1} w_n) \quad \cdots (1)$$

This can be approximately estimated by the following formula.

$$\begin{aligned} PW^* &= \max_{s_1, s_2, \cdots, s_{n-1}} P(w_1 s_1 w_2 s_2 \cdots w_{n-1} s_{n-1} w_n) \\ &\approx \max_{s_1, s_2, \cdots, s_{n-1}} \{P(pos_1) P(s_1 | pos_1) P(s_2, pos_2 | s_1, pos_1) \\ &\cdots P(s_{n-1}, pos_{n-1} | s_{n-2}, pos_{n-2}) P(s_n, pos_n | s_{n-1}, pos_{n-1})\} \end{aligned} \quad \cdots (2)$$

Since $P(pos_1)$ is a constant value here, the above formula can be simplified as

$$\begin{aligned} PW^* &= \max_{s_1, s_2, \cdots, s_{n-1}} P(w_1 s_1 w_2 s_2 \cdots w_{n-1} s_{n-1} w_n) \\ &\approx \max_{s_1, s_2, \cdots, s_{n-1}} \{P(s_1 | pos_1) P(s_2, pos_2 | s_1, pos_1) \\ &\cdots P(s_{n-1}, pos_{n-1} | s_{n-2}, pos_{n-2}) P(s_n, pos_n | s_{n-1}, pos_{n-1})\} \end{aligned} \quad \cdots (3)$$

To calculate the above formula, five kinds of probabilities should be estimated from our training corpus. These are described below.

1) The probability that a POS is the POS of a singleton PW; namely,

$$P(s = I | pos = pos\_i) = \frac{C(s = I, pos = pos\_i)}{C(s = I, pos = pos\_i) + C(s = B_1, pos = pos\_i)}$$

2) The probability that a POS is the POS of the first lexical word in a PW, assuming that this PW contains at least two lexical words; namely,

$$P(s = B_1 | pos = pos\_i) = \frac{C(s = B_1, pos = pos\_i)}{C(s = I, pos = pos\_i) + C(s = B_1, pos = pos\_i)}$$

3) The transition probability from position B_1 to position B_2 within a PW.

$$P(s=B_2, pos=pos\_j \mid s_{prev}=B_1, pos_{prev}=pos\_i) = \frac{C(s_{prev}=B_1, pos_{prev}=pos\_i, s=B_2, pos=pos\_j)}{C(s_{prev}=B_1, pos_{prev}=pos\_i)}$$

4) The transition probability distribution from position B_2 or position M to position M.

$$P(s=M, pos=pos\_j \mid s_{prev}=B_2 orM, pos_{prev}=pos\_i) = \frac{C(s_{prev}=B_2 orM, pos_{prev}=pos\_i, s=M, pos=pos\_j)}{C(s_{prev}=B_2 orM, pos_{prev}=pos\_i)}$$

5) The jump probability in a PW boundary.

$$P_{jump}(pos=pos\_j \mid pos_{prev}=pos\_i) = P(s=B_1 orI, pos=pos\_j \mid s_{prev}=B_2 orMorI, pos_{prev}=pos\_i) = \frac{C(s_{prev}=B_2 orMorI, pos_{prev}=pos\_i, s=B_1 orI, pos=pos\_j)}{C(pos_{prev}=pos\_i, pos=pos\_j)}$$

The global PW grouping probabilities of a sentence with various possible grouping paths will be calculated using the above five probabilities by the dynamic programming approach. The path with the biggest probability will be treated as the optimum PW grouping result.

4. Experiment results and discussion

4.1. Test set

An independent test corpus, also selected from the People's Daily Corpus 1998, was used in the experiments reported in this paper. This test set contained 400 sentences with an average number of Chinese characters per sentence of about 37 and average number of lexical words per sentence of about 23. These figures are consistent with actual cases. The prosody structure was labeled by a well-trained annotator from the text and then modified by listening to the speech corpus recorded by a female graduate student majoring in Chinese literature. Finally, the test set was annotated with 5113 PW boundaries.

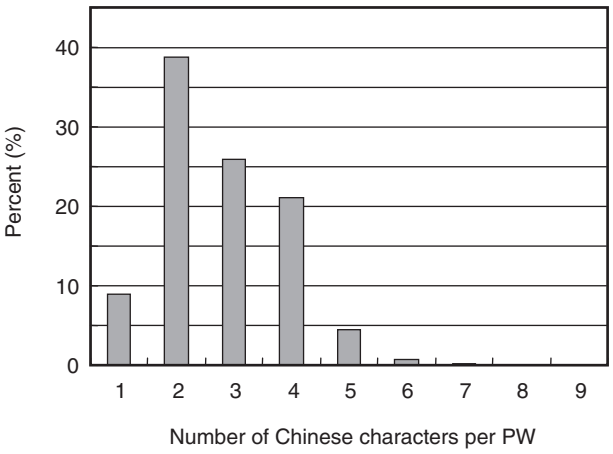

Figure 3
PW length histogram.

A length histogram of PWs in the 400-sentence test set is shown in **Figure 3**. The average number of Chinese characters per PW in the test set was about 2.8. Our data was quite different from that in Reference 3), where the authors assumed that a PW is primarily composed of disyllables or trisyllables. However, many relative long units, in particular, those with two lexical words such as "宽阔明亮" and "筹措资金" were annotated as PWs in our test set as a result of listening. An example of a transcribed sentence in our test set is:

400. 他们/r | 着意/d 揣摩/v | 专家/n 意图/n ，/w ||| 反复/d 征求/v | 专家/n 意见/n ，/w ||| 因为/c || 只有/c 专家/n | 满意/v 了/y ，/w ||| 作品/n || 才/c 有/v 希望/n | 获奖/v 。/w@

4.2. Results and discussion

Precision and recall statistics were calculated to evaluate the performance of PW grouping for these results. Experimental results for PW grouping by our global probability estimation method are given in **Table 1**, which also gives the results when a module for automatic word segmentation and POS tagging was applied. The results were better than those for our previous binary prosodic tree method.[4)]

As described in Section 4.1, many relatively long units, in particular, those with two lexical

Table 1
Experiment results.

Methods	Precision	Recall rate
New method with transcribed word segmentation & POS tagging	89.56%	84.37%
New method with automatic word segmentation & POS tagging	87.52%	82.84%
Binary prosodic tree method (previous method)	85.91%	79.38%

Table 2
Comparison of resource requirements.

Methods	ROM cost	RAM cost
Global PW probability estimation method (new method)	1.0 MB	1.0 MB
Binary prosodic tree method (previous method)	35 MB	30 MB

words (two feet), such as "宽阔明亮" and "筹措资金", were annotated as PWs through listening. Therefore, the PW grouping module must also resolve these kinds of PW groupings. An annotator was asked to check all of the wrong predicted prosodic boundaries. About half of them arose from PW boundary insertion errors that occurred within those relative long PWs, for example, "宽阔 | 明亮" and "这 位 | 导演". Although perception experiments show that it is better to group these kinds of PWs, it is also acceptable not to group them in synthesized speech. In fact, sometimes people also regard them as two different PWs in speech for emphasis purposes or for poetic style.

The resource requirements for our new method and for the previous binary prosodic tree method are compared in **Table 2**. From this table, we can see that both the read-only memory (ROM) and random access memory (RAM) costs of the new method are much lower than those of the previous method. Lower ROM and RAM costs for the TTS system are very important for some applications such as call center applications and embedded applications because call center applications need to support multichannel processes and embedded devices have limited resources.

5. Conclusion

In this paper, we described a global probability estimation method based on the positional types of lexical words according to their positions in the PWs that they belong to. The optimum PW grouping path of a sentence can be obtained through a dynamic programming approach. The experimental results are very promising. They are better than those for our previous binary prosodic tree method in terms of both accuracy and memory cost.

References

1) A. Li and M. Lin: Speech corpus of Chinese discourse and the phonetic research. International Conference on Spoken Language Processing, Beijing, 2000, Vol. 4, pp. 13–18.
2) Q. Guo, E. Xun, and N. Katae: Prosody word grouping in Mandarin TTS system. International Symposium on Chinese Spoken Language Processing, Singapore, 2006, Vol. 2, pp. 181–190.
3) M. Chu and Y. Qian: Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts. *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 6, No. 1, pp. 61–82 (2001).
4) Q. Guo, N. Katae, H. Yu, and H. Iwamida: High Quality Prosody Generation in a Text-to-speech System. (in Chinese). *Journal of Chinese Information Processing*, Vol. 22, No. 2, pp. 110–115 (2008).

Qing Guo
Fujitsu Research and Development Center Co., Ltd.
Dr. Guo received a Ph.D. degree in Computer Science and Application from Tsinghua University, Beijing, China in 1999. He joined Fujitsu Research and Development Center Co., Ltd., Beijing, China in 2004 and has been engaged in research and development of speech synthesis technologies.

Jie Zhang
Fujitsu Research and Development Center Co., Ltd.
Ms. Zhang received an M.S. degree in Chinese Language and Literature from Peking University, Beijing, China in 2006. She joined Fujitsu Research and Development Center Co., Ltd., Beijing, China in 2006 and has been engaged in research and development of natural language processing.

Nobuyuki Katae
Fujitsu Laboratories Ltd.
Mr. Katae received B.S. and M.S. degrees in Design from Kyushu Institute of Design, Fukuoka, Japan in 1989 and 1991, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1991 and has been engaged in research and development of speech synthesis technologies. He is a member of the Acoustical Society of Japan.

Hao Yu
Fujitsu Research and Development Center Co., Ltd.
Dr. Yu received a Ph.D. degree in Electrical Engineering from Harbin Institute of technology, Harbin, China in 1998. He joined Fujitsu Research and Development Center Co., Ltd., Beijing, China in 2003 and has been engaged in research and development of natural language processing.