Prosody Generation in a Chinese TTS System Based on a Hierarchical Word Prosody Template Tree

Chung-Hsien Wu and Jau-Hung Chen

Institute of Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.
E-mail: {chwu, chenjh}@server2.iie.ncku.edu.tw

Abstract

In this paper, a prosody generation method based on a hierarchical word prosody template tree for the generation of prosodic information in a sentence is proposed. The hierarchical word prosody template tree is established from a large continuous speech database according to the linguistic features: tone combination, word length, part of speech (POS) of the word, and word position in a sentence. This template tree stores the prosodic features including pitch contour, energy contour, and syllable duration of a word for possible combinations of linguistic features. In the inside test, the prosody of the synthesized speech resembles that of the original speech for a typical sentence.

1. Introduction

In recent years, text-to-speech systems have been able to generate highly intelligible synthesized speech. However, further improvement of synthesized speech is expected with respect to the prosodic information. Two general approaches have been proposed for generation of the prosodic information: the rule-based approach (D. H. Klatt 1987, L. S. Lee 1989) and the data-driven approach (C. A. Moor 1995, H. S. Hwang 1995). However, the rule-based and data-driven approaches use small number of prosodic patterns to represent the diverse prosody. Consequently, the prosody generated by rules or neural networks is an average version of all the original prosody in the same linguistic conditions. These approaches cannot give the best result even though the sentence (or the linguistic features) are identical to a sentence (or the linguistic features) in the training database. Therefore, a more detailed and accurate description between prosodic and linguistic features of a sentence is desired to achieve a better synthesized result.

In this paper, a word prosody template tree recording all the relationship between the linguistic features and the word prosodic templates in the speech database is established. Each word prosodic template contains the syllable duration, energy contour and pitch contour of the word. For each word in a sentence/phrase, the word position is first determined and used to traverse the template tree. Word length is then used to find the tone combination subtree. Finally, tone combination for the word is used to retrieve the word prosody template.

2. Construction of Word Prosody Template Tree

In the Chinese TTS system, some linguistic features are relevant to word prosodic information. They are tone combination, word length, POS of the word, and word position in a sentence. These features are discussed in more detail in the following.

(1) Tone combination: A word with length n consists of n syllable(s) in which each syllable has a tone. However, the neural tone generally appears at the end of a word. As a result, there are $4^{n-1} \cdot 5$ tone combinations for an *n*-syllable word.

(2) Word length: The intonational or prosodic relationship between syllables within a word is more obvious than that between two words. Therefore, word length of an *n*-syllable word is used to choose its corresponding word prosodic patterns with word length *n*.

(3) POS of the word: POS is also an important linguistic feature to determine the word prosody. In this paper, POS is divided into 21 categories. The distance between two categories is defined by the distance of their corresponding average prosodic patterns in the training database, i.e., word pitch contour, word energy contour, and syllable duration in the word. This distance is then normalized to lie between 0 and 1. Therefore, a POS distance table was established.

(4) Word position in a sentence: A word position ratio is defined as the order of the word position in the sentence divided by the number of words in the sentence.

Using the above linguistic features, a word prosody template tree is constructed. The linguistic features of a word, i.e., tone combination, word length, POS of the word and word position in a sentence, are associated with a set of prosodic patterns, i.e., syllable duration, energy contour, and pitch contour. To establish the word prosody template tree, a continuous speech database established by the Telecommunication Laboratories, Chunghwa Telecom Co., Taiwan, containing 655 reading utterances was used. The speech signals were digitized by a 16-bit A/D converter at a 20-kHz sampling rate. The syllable segmentation and phonetic labels were manually done. A total number of 38907 syllables and their phonetic labels were obtained. Using the text analysis, 9698 reference words (including 2-, 3-, and 4-syllable words) and their corresponding word prosodic patterns were obtained.

The structure of the word prosody template tree is shown in Fig. 1. There are three levels: word position level, word length level and tone combination level. In the word position level, an input word pattern is classified into one of the following three categories: the beginning part of a sentence (BOS), the middle part of a sentence (MOS), and the end part of a sentence (EOS). Furthermore, each category has four branches according to word length. As illustrated in the word length level, they are monosyllable words, 2-syllable words, 3-syllable words, and 4-syllable words. Finally, the tone information of a word is described in the tone combination level. This level contains five groups according to their ending tones. The reason for this grouping configuration is to link the word prosody correlation between adjacent words. For each tone combination, a word prosody template is established to store the prosodic features: pitch contour, energy contour, and syllable duration. Besides, the POS and the word position ratio of the word are also stored.

3. Generation of Word Prosody Templates

The generation of the word prosody templates is shown in Fig. 2. An input sentence/phrase is first decomposed into a sequence of words by a word segmentation parser. Each word contains linguistic features including tones, word length, POS, and word position in a sentence. In our system, a word with length more than four is further divided into combinations of 1- to 4-syllable words. The deep first search algorithm (G. Chartrand 1993) is

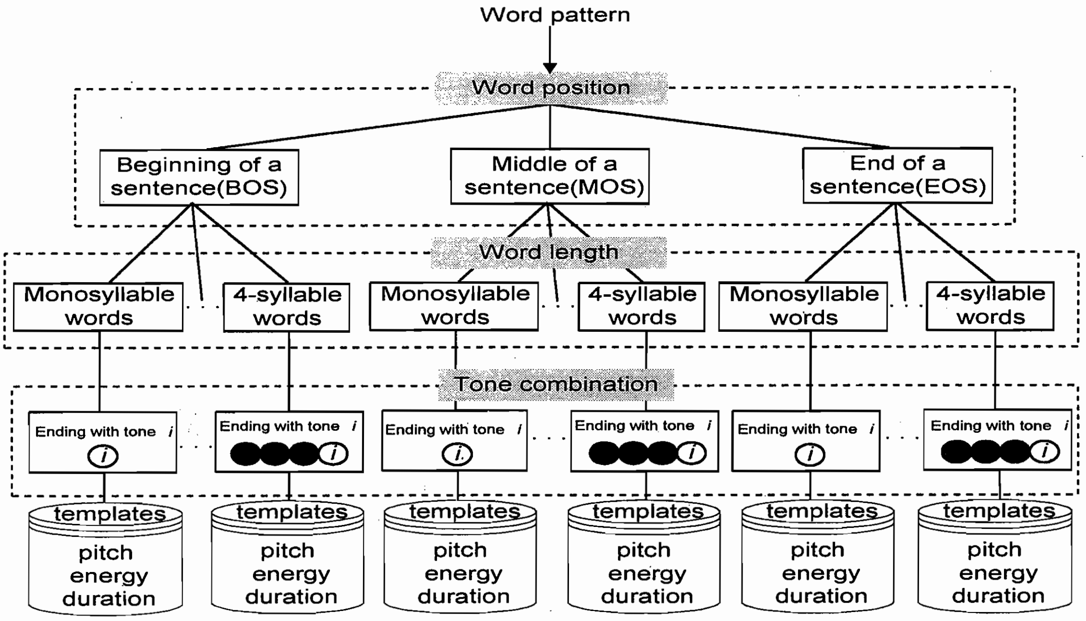

Fig. 1. Structure of the hierarchical word prosody template tree for word templates ending with tone *i*.

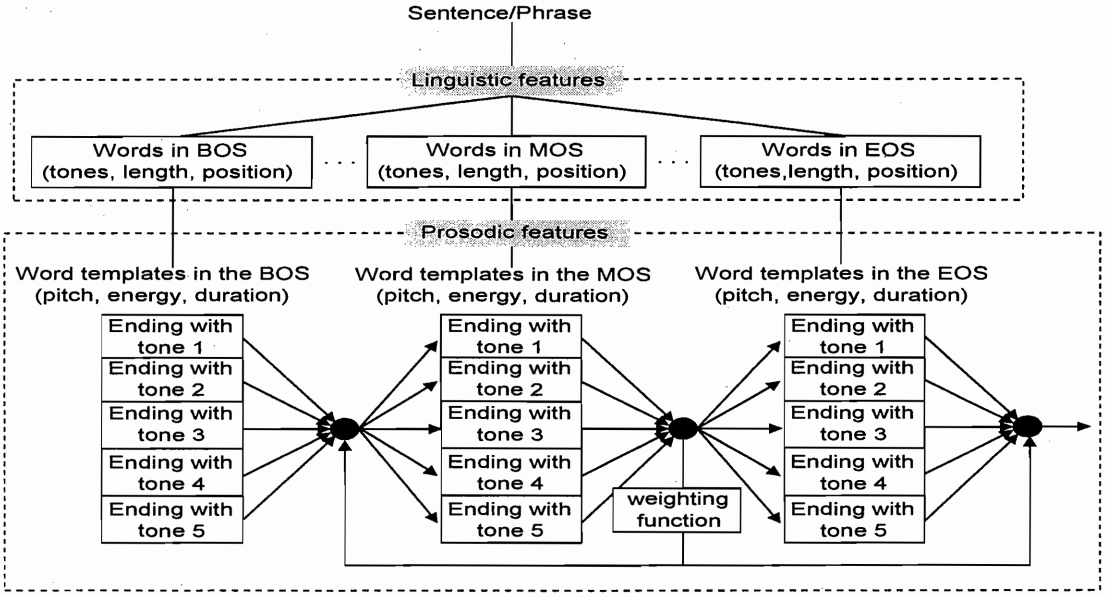
Sentence/Phrase
Linguistic features
Words in BOS (tones, length, position)
Words in MOS (tones, length, position)
Words in EOS (tones,length, position)
Prosodic features
Word templates in the BOS (pitch, energy, duration)
Word templates in the MOS (pitch, energy, duration)
Word templates in the EOS (pitch, energy, duration)
Ending with tone 1
Ending with tone 2
Ending with tone 3
Ending with tone 4
Ending with tone 5
weighting function

Fig. 2. Generation of the word prosody templates for a sentence/phrase.

employed to traverse the word prosody template tree. For each input word in a sentence, the word position and word length were first used to reach the word length level. Second, the tone combination is used to locate the corresponding word prosody template(s). There are two conditions of the target template in the tree: reachable and unreachable. They are discussed as follows.

(1) Reachable: If a target template can be reached and there is only one template found then output it. Otherwise, one of the template candidates is selected by calculating the linguistic distance between the input word and the reference words in the tree. The prosodic template corresponding to the reference word with the minimum linguistic distance is chosen as the output using the following distance estimation method.

$$j^{*}=\min_{j}\{d_T(T_I,T_j)+d_C(C_I,C_j)+d_P(P_I,P_j)\},\quad j=1,\cdots,J \qquad (1)$$

where J is the total number of words in a sentence corresponding to a given word position, word length, and tone combination. $d_T(T_I,T_j)$ represents the pitch distance between the average pitch of the input word T_I and that of the reference word T_j in the tree. $d_C(C_I,C_j)$ represents the linguistic distance between the POS of the input word C_I and that of the reference word C_j in the tree. $d_P(P_I,P_j)$ represents the absolute distance between word position ratio of the input word P_I and that of the reference word P_j.

(2) Unreachable: When a target template is unreachable, that means the linguistic feature of the input word does not appear in the speech database. There is no corresponding prosodic pattern in the template tree. However, we should find a suitable prosodic pattern to correspond to the input linguistic feature. In this case, the other two subtrees in the word length level are traversed in the following order according to the current word position: (A) If the current word position is BOS then traverse MOS subtree followed by EOS subtree. (B) If the current word position is MOS then traverse BOS subtree followed by EOS subtree. (C) If the current word position is EOS then traverse MOS subtree followed by BOS subtree.

4. Results

Fig. 3 illustrates an example of pitch contours of the original speech and synthesized speech. The first two panels display the waveform and the pitch contour of the original speech selected from the speech database. The last panel shows the corresponding pitch contours of the synthesized speech. By examination of the pitch contours, the synthesized pitch contours generated from the hierarchical word prosody template tree resemble their original counterparts. The results of listening test also confirm the good performance of this scheme.

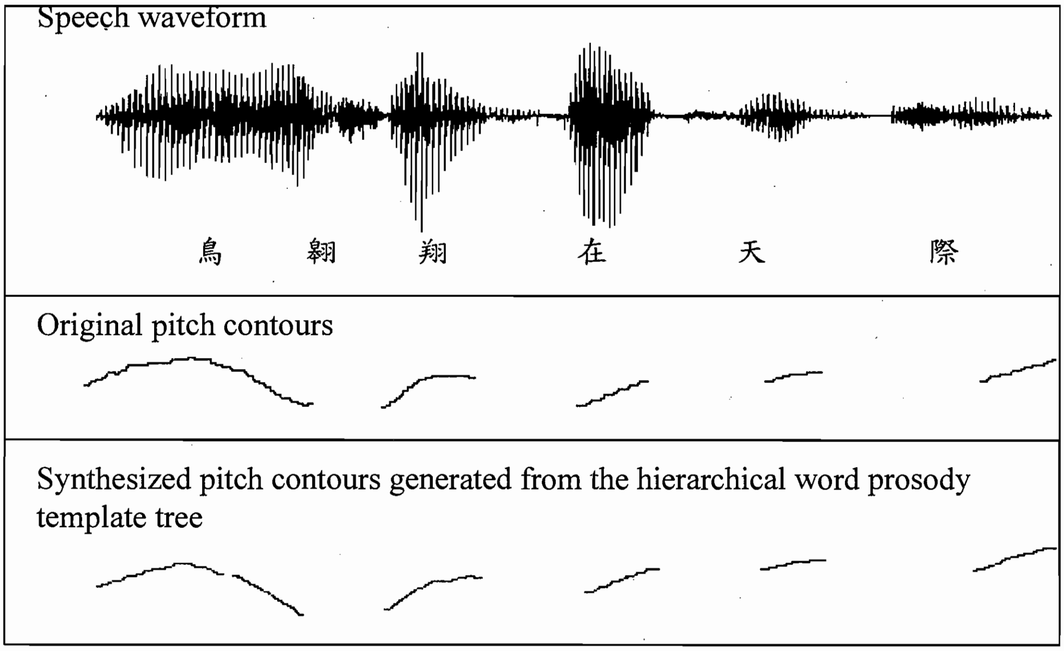

Fig. 3. An example of pitch contours of the original speech and synthesized speech.

5. Conclusions

In this paper, the construction and generation of prosodic information for Chinese text-to-speech conversion has been proposed to enhance the conventional rule-based approach. The prosodic information including pitch contour, energy contour, and syllable duration, was stored in a hierarchical word prosody template tree generated from a large speech database. Appropriate word prosodic templates in a sentence are selected from the tree according to the linguistic features. Evaluation by means of listening tests has confirmed the good performance of this scheme.

References

Klatt, D. H., "Review of text-to-speech conversion for English," J. Acoust. Soc. Amer., vol.82, No.3, 1987, pp.737-793.

Lee L. S., C. Y. Tseng, and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," IEEE Trans. ASSP, Vol.37, 1989, pp. 1309-1320

Moor C. A., J. F. Cohn and G. S. Katz, "Quantitative description and differentiation of fundamental frequency contours," Computer Speech and Language, Vol.8, 1994, pp.385-404.

Hwang S. H. and S. H. Chen, "A prosodic model of Mandarin speech and its application to pitch level generation for text-to-speech," in Proc. ICASSP, 1995, pp. 616-619.

Chartrand G. and O. R. Oellermann, *Applied and algorithmic graph theory*, McGraw-Hill, Inc., 1993, p.74.