Template-Driven Generation of Prosodic Information for Chinese Concatenative Synthesis

Chung-Hsien Wu and Jau-Hung Chen
Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.

ABSTRACT

In this paper, a template-driven generation of prosodic information is proposed for Chinese text-to-speech conversion. A set of monosyllable-based synthesis units is selected from a large continuous speech database. The speech database is employed to establish a word-prosody-based template tree 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, average energy, and syllable duration of a word for possible combinations of linguistic features. Two modules for sentence intonation and template selection are proposed to generate the target prosodic templates. The experimental results for the TTS conversion system showed that synthesized prosodic features quite resembled their original counterparts for most syllables in the inside test. Evaluation by subjective experiments also confirmed the satisfactory performance of these approaches.

1. INTRODUCTION

In past years, many studies have focused on TTS systems for different languages [1],[2]. Also, TTS systems and synthesis technology for the Chinese language have been developed in the last two decades [2]-[4]. In concatenative speech synthesis, rule-based approach has been used for prosody modification [1], [2]. These phonological rules are invoked to imitate the pronunciation of humans. The derivation of phonological rules, however, is laborsome, time wasting and tedious. Furthermore, because many various linguistic features interactively affect the phonological characteristics, it is difficult to collect appropriate and complete rules to describe the prosody diversity. Consequently, a novel approach using neural networks has been investigated for automatic learning of prosodic information [3]. However, the network can become trapped in a local minimum of the error function, thus arriving at an unacceptable solution when a better one exists.

This paper proposes a Chinese text-to-speech conversion system, which focuses on the generation of prosodic information. In this approach, the word is chosen as the unit for prosody modification because word is the basic rhythmical pronunciation unit. The intonational or prosodic relationship between syllables within a word is more obvious than that between two words. Furthermore, it appears that the prosodic properties of a Chinese word is generally affected by the tone combination, word length, part of speech of the word, and word position in a sentence. Therefore, a word prosody template tree recording the relationship between the linguistic features and the word prosody templates in the speech database is established. Each word prosody template contains the syllable duration, average energy and pitch contour of the word. For each word in a sentence/phrase, word length is first determined and used to traverse the template tree. Tone combination is then used to retrieve the stored templates. Finally, a sentence intonation module and a template selection module are proposed to select the target prosodic templates. The time-domain/waveform pitch-synchronous overlap-and-add (PSOLA) method is employed for the modification of prosodic information [5].

2. TEMPLATE-DRIVEN PROSODY GENERATION

An important characteristic of Mandarin Chinese is that it is a tonal language based on monosyllables. Each syllable can be phonetically decomposed into an initial part followed by a final part. Five basic tones are the high-level tone (Tone 1), the mid-rising tone (Tone 2), the midfalling-rising tone (Tone 3), the high-falling tone (Tone 4), and the neutral tone (Tone 5). From the viewpoint of Chinese phonology, the total number of phonologically allowed syllables in Mandarin speech is only about 1300. Therefore, a syllable is a linguistically appealing synthesis unit in a Chinese TTS system. In this paper, the set of 1313 tonal monosyllables is adopted as the basic set of synthesis units, which was selected from a large continuous speech database. For each speech unit in the speech database, pitch-mark labeling is automatically estimated by the autocorrelation method. Besides, quantitative description of the pitch contours is expressed by orthonormal expansion using discrete Legendre polynomials [6]. That is, a pitch contour can be represented by a four-dimensional vector (a_0, a_1, a_2, a_3).

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

1. Tone combination of the word: In Mandarin Chinese, the word is the basic comprehensive pronunciation unit. The intonation of a word is primarily reflected in F_0 contours, which adequately represent the lexical tones. Moreover, the same tone adjacent to different tones results in different F_0 contours, which might vary in shapes, slopes and means. A word with length n consists of n syllable(s) in which each syllable has a lexical 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: In Mandarin speech, some of the prosodic characteristics of the phrase, such as final lengthening effect, are presented in the word. Therefore, for an *n*-syllable word, the word length, *n*, is used to obtain the corresponding prosodic templates.
3. POS of the word: A word might differ from itself in POS, which carries various prosodic information. In this paper, POS is divided into 18 categories to capture the variations of prosodic features.
4. Word position in a phrase: In general, the pitch contour and energy contour in a phrase will follow an intonation pattern. For example, the F_0 contour and the energy contour will decline in a declarative sentence.

In this paper, a word prosody template tree is constructed based on the above linguistic features. To establish the word prosody template tree, a large continuous speech database was used. Using the text analysis module, 21348 reference words (including 1- to 4-syllable words) and their corresponding prosodic patterns were obtained. The number of word patterns in the database and the tone combinations are shown in Table 1.

Table 1. Distribution of word prosody patterns in the database

Word length	Tone combinations	No. of word templates
1	5	9839
2	20	9703
3	80	1226
4	320	580

2.1 Structure of the word prosody template tree

The structure of the word prosody template tree is shown in Fig. 1. This template tree contains two levels: word-length level and tone-combination level. In the word-length level, it includes monosyllable words, 2-syllable words, 3-syllable words, and 4-syllable words. Each child of the node in the word-length level is further categorized by tone combination in the tone-combination level. For each tone combination, the word prosody templates are established to store the prosodic and linguistic features. For each syllable in the word, the stored prosodic features are as follows:

- Pitch vector: The pitch vector represented by (a_0, a_1, a_2, a_3) is stored.
- Energy: The average energies of the initial and the final parts are stored, respectively.
- Duration: The duration of the initial and the final parts are stored, respectively.

And the stored linguistic features include

- POS of the word.
- The group numbers of the initial and the final parts
- The codes of the initial and the final parts.
- Big5 codes of the word.

In general, the F_0 contour will decline in a declarative sentence. Consequently, the word prosody templates are increasingly sorted according to their average pitch periods (a_0's) for each tone combination. The merit of this alignment is that the prosodic effects of word position in a phrase are implicitly built and can be retrieved easily.

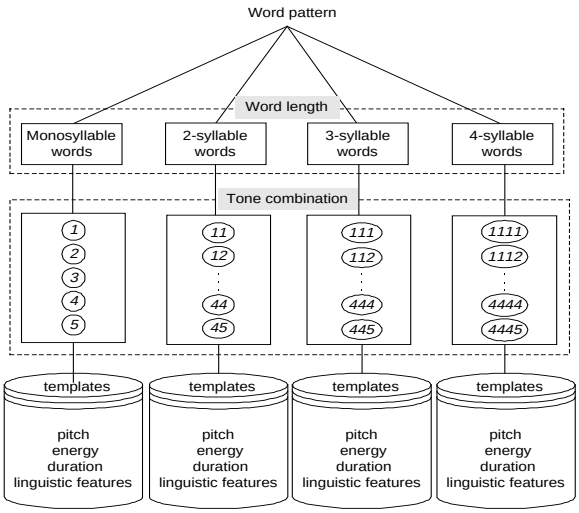

Fig. 1. Structure of the word prosody template tree.

2.2 Generation of Word Prosody Templates

The generation of the word prosody templates is shown in Fig. 2. An input phrase/sentence is first decomposed into a word sequence by the word segmentation module in text analysis. In the word sequence, each word is labeled with linguistic features including phonemes, tones, word length, POS, and word position in a sentence. To obtain the prosodic features from the word template tree, a sentence intonation module is used to compute the target pitch period for the first syllable of the word. The linguistic features of each word and its target pitch period are then fed to template selection module. Based on the target pitch period, this module uses a cost function to estimate the distance of linguistic features between the input word and the one in the template tree. These two modules are described in detail as follows.

1. Sentence Intonation Module

The sentence intonation module provides a global F_0 contour for the synthesized speech. By inspecting the global F_0 contours of the phrases in the speech database, we found that the speaker pronounced high F_0 at the beginning and low F_0 at the end, that is, global pitch-period contours are gradually increasing. Also, the pitch-period contours of the words are raised one by one which constitute the pitch-period contour of a phrase. In the word level, the word intonation and precise pitch-period variation are captured in every word. Since the word pitch-period contours have been stored in the word template tree, the key role of this module is to provide the word intonation. For the *m*th word of the word sequence in Fig. 2, the triplet (*t*,*l*,*L*) is the input of this module in which *t* is the tone of the first syllable in the word and *l* is the first syllable's position in the sentence with *L* syllables. The target pitch period of the first syllable of the *m*th word is obtained by the following equation:

$$TP_m(t,l,L) = p_t \cdot r(l,L) \quad (1)$$

where P_t represents the average pitch period for each tone in the speech database, $1 \le t \le 5$. The function *r(l,L)* is called a *ratio function*, $0< r(l,L)<2$ for $1 \le l \le L$, which is defined as

$$r(l,L) = r_0 + \left(\frac{2}{1+v^{\frac{L+1}{2}-l}} \right) \cdot (1-r_0) \quad (2)$$

In this equation, r_0 is the offset of the ratio function and is referred to as minimal ratio with $0<r_0<1$. The variable *v* represents the changing rate of this function. In order to generate different values of this function, *v* is assigned a random number between 1.5 and 3. Also, it can be seen that the syllable/word in the beginning and the end of the sentence obtain ratios smaller and larger than 1, respectively. And that in the middle of the sentence obtains a ratio approaching to 1. Therefore, Eq. (9) will generate the target sentence intonation represented by a rising pitch contour centered with an adequate value of average pitch period.

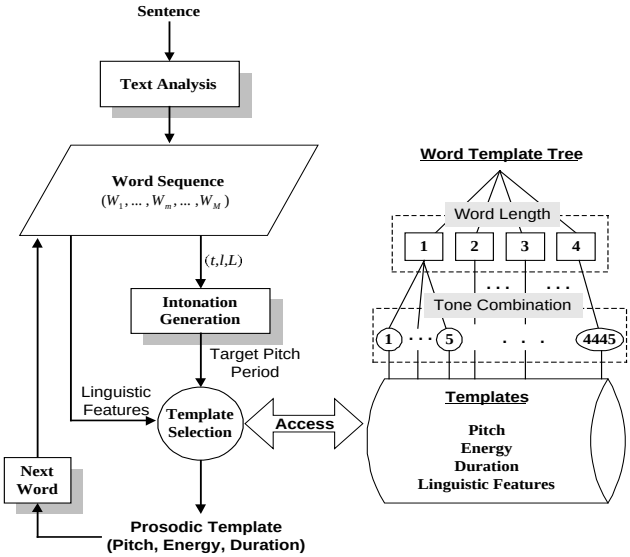

Fig. 2. Generation of the word prosody templates for a sentence.

2. Template Selection Module

For the *m*th word W_m of the word sequence, the template selection module first traverses the word template tree according to the word length and the tone combination. Second, the target pitch period TP_m obtained from the sentence intonation module is used to choose a small set of word prosody template(s) from the stored templates. The set of word template candidates is constructed by selecting the templates with average pitch period close to TP_m. It is obtained as follows:

$$WT = \{ t_i \mid |TP_m - a_{0_i}| < TH, 1 \le i \le \text{number of all templates} \} \quad (3)$$

where a_{0_i} is the component of the pitch vector of the first syllable in the *i*th word template. *TH* is a tolerance for the variety of average pitch period and set to 10 samples in our system. Third, a cost function is used to compute the distance of the linguistic features between the input word and the one in the set *WT*. The template cost function is expressed as

$$WTC(W_m, t_n) = D_0(POS_m, POS_n) + \sum_{i=1}^{|W_m|} \{ D_0(IG_{mi}, IG_{ni}) + D_0(FG_{mi}, FG_{ni}) + D_0(I_{mi}, I_{ni}) + D_0(F_{mi}, F_{ni}) + D_0(Big5_{mi}, Big5_{ni}) \} \quad (4)$$

The notations in this equation are described in the following.

- t_n: A word template in *WT*.
- $D_0(x,y)$ A distance measure between *x* and *y* defined as
$$D_0(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \ne y \end{cases} \quad (5)$$
- *POS*: POS of the word.
- $|W_m|$: Word length of W_m.
- IG_{mi} and FG_{mi}: Group numbers of the initial and the final parts of the *i*th syllable in W_m, respectively.
- I_{mi} and F_{mi}: Codes of the initial and the final parts of the *i*th syllable in W_m, respectively.
- $Big5_{mi}$: Big5 code of the *i*th syllable in W_m.
- The symbols with index *n* are linguistic features for t_n.

Finally, the prosodic features of the word template with the minimal cost are retrieved as the target prosodic features for W_m, $1 \le m \le M$.

3. EXPERIMENTS AND RESULTS

In our system, a continuous speech database established by the Telecommunication Labs., Chunghwa Tele-communication Co., Taiwan, containing 655 reading utterances was used to construct the word template tree. 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.

3.1 Average Pitch Periods of the Five Tones

In the speech database, the distribution of the pitch periods (in sample) of the five tones is shown in Table 2. For the male speaker, panel (a) indicates that Tone 1 and Tone 4 have smaller values of average pitch periods (higher mean F_0) while Tone 3 and Tone 5 have larger values of average pitch periods (lower mean F_0). Among the five tones, Tone 2 has a medium value of average pitch period. For the standard deviation listed in the third column, it can be seen that the five tones have less difference between each other. The precise description of the distribution of pitch period is shown in panel (b). The shadow area in each row indicates the dominative range of pitch period for each tone. Although the first four tones have the same range in the shadow areas, the variations of mean F_0 value are not all the same and listed as follows: 77Hz (i.e., 20000/110 – 20000/190), 59Hz, 46Hz, 77Hz, 38Hz for Tone 1 to Tone 5, respectively. It implies that Tone 1 and Tone 4 were pronounced by a wider pitch variation than the others. The neural tone has the least pitch variation (38Hz) which is only half of that of Tone 1 and Tone 4.

3.2 Inside Test

In this experiment, a training text was chosen for the inside test of the proposed approach. Fig. 3 shows an example of the original speech and synthesized prosodic parameter sequences of the mean pitch period, average energy, initial part duration, and final part duration. For the mean pitch period of each syllable in the test text, panel (a) plots the the synthesized and the original contours. It can be seen that the two contours are very alike for most syllables. The results of mean energy, initial part duration, and final part duration are shown in panel (b), (c), and (d), respectively. It can be seen that the contours of some syllabes match quite well with their counterparts. However, obvious deviation occur at some syllables. This is because this word is not in our word lexicon and it is decomposed into two monosyllable words. On the other hand, initial part duration is more syllable dependent and not with large duration variation compared to the other three features. Among them, the mean pitch period is less syllable dependent.

Table 2. Distribution of the pitch periods (in sample) of the five tones. Panel (a) displays the average pitch periods and the standard deviations. Panel (b) lists the range of the pitch periods and corresponding percentages (%)

Tone	Average Pitch Period	Standard Deviation
1	141	23
2	168	24
3	185	28
4	149	26
5	179	27

(a)

Tone	Range of the pitch periods (in sample) and corresponding percentage (%)							
	110-	110~ 130	130~ 150	150~ 170	170~ 190	190~ 210	210~ 230	230+
1	7.4	26.4	31.2	22.5	10.8	1.7	0	0
2	0.4	4.8	16.8	29.0	29.5	16.0	3.3	0.2
3	1.4	1.3	4.7	19.5	31.7	25.3	10.3	5.8
4	4.7	19.0	28.0	25.6	15.1	6.0	1.4	0.2
5	0.8	4.3	8.0	18.7	31.8	25.3	8.2	2.9

(b)

4. CONCLUSIONS

In this paper, the approach to the generation of prosodic information has been proposed for a Chinese Text-to-Speech system using a large speech database. The prosodic information was stored in a word prosody template tree along with the linguistic features. The proposed sentence intonation module generated a sequence of target pitch periods by means of a ratio function and average pitch periods. The template selection module selected appropriate prosodic templates from the tree according to the target pitch periods and the linguistic features. Experimental results showed that the synthesized pitch contours quite resembled their original counterparts. The result of listening test confirmed that the synthesized speech is highly intelligible and natural.

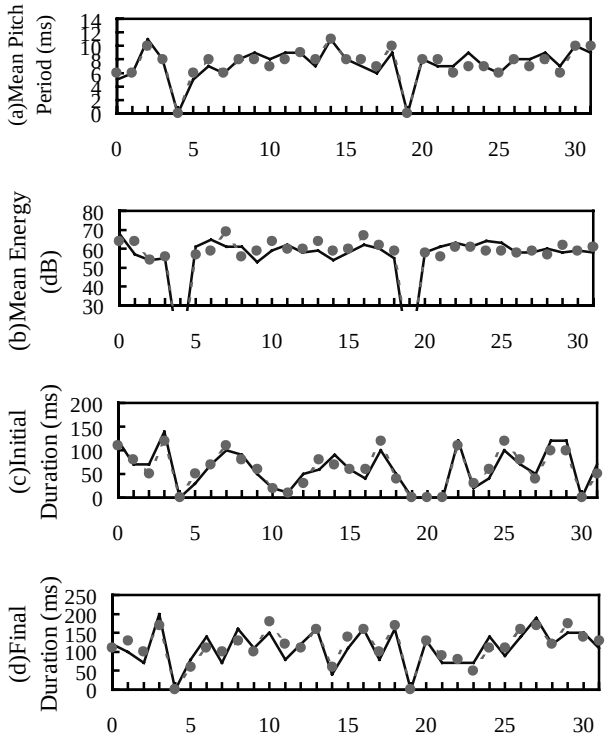

Fig. 3. Example of the original (solid lines) and the synthesized (dotted lines) prosodic parameter sequences of: (a) mean pitch period, (b) average energy (c) initial part duration, and (d) final part duration. The X-axis represents the syllable positions corresponding to the Chinese characters.

5. REFERENCES

[1] D. H. Klatt, "Review of text-to-speech conversion for English." *J.Acoust. Soc. Amer*, **82**(3), pp. 737-793, 1987.

[2] L. S. Lee, C. Y. Tseng and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," *IEEE Trans. Acoust, Speech, Signal Processing*, **37**(9), pp. 1309-1319, 1989.

[3] S. H. Chen, S. H. Hwang and Y. R. Wang, "An RNN-based prosodic information Synthesizer for Mandarin text-to-speech," *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 3, pp. 226-239, 1998.

[4] C. L. Shih and R. Sproat, "Issues in text-to-speech conversion for Mandarin," in Computational Linguistics and Chinese Language Processing, vol.1, pp.37-86, 1996.

[5] F. J. Charpentier and M. G. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," *Proc*. ICASSP, pp. 2015-2020, 1986.

[6] S. H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Trans. on Commun.*, **38**, pp. 1317-1320, 1990.