Joint Conference of 1996 International Computer Symposium
December 19~21, Kaohsiung, Taiwan, R.O.C.

Synthesis Unit Selection and Word Prosody Adjustment in a Chinese Text-to-Speech System *

Chung-Hsien Wu, Jau-Hung Chen and Tso-Chih Lin

Institute of Information Engineering
National Cheng Kung University, Tainan, Taiwan, R.O.C.
chwu, chenjh@server2.iie.ncku.edu.tw

Abstract

This paper proposes a Chinese text-to-speech conversion system which focuses on the selection of synthesis unit and word prosodic pattern. In order to take into account both prosodic and phonetic information between concatenated syllables, synthesis units are selected from a large continuous speech database. On the other hand, since word is the basic rhythmical pronunciation unit, a word-based prosody modification approach is proposed. From observation, 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. Consequently, a word prosody database is established based on above linguistic features. Each word pattern in the word prosody database contains linguistic features and prosodic patterns of the word. Moreover, the cubic spline curve fitting method is adopted to describe the prosodic pattern. A word prosody selection method was proposed to select appropriate word prosodic pattern from the word prosody database. According to the result of listening test, an adequate sequence of synthesis units and word prosodic patterns was obtained.

1. Introduction

In concatenative speech synthesis, the choice of synthesis units to be concatenated plays a prominent role of synthesizing intelligible and natural high quality speech. In past years, many kinds of synthesis units have been proposed [1]. The phonemes have been adopted as the basic synthesis units. Such units take advantage of small storage. However, the accuracy of coarticulation and the spectral discontinuity between adjacent units need to be improved. Consequently, longer synthesis units, such as diphone, demi-syllable, syllable, triphone and polyphone, are appropriately incorporated to reduce the effect of spectral distortion [1][7]. Recently, the approaches of unit selection from a large speech database or using non-uniform units [2][3] have been appreciated and proved to obtain natural and high quality speech. This approach employes some criteria or a cost function to select an appropriate synthesis segment sequence.

On the other hand, rule-based approach have been used for prosody modification [1], [4]-[5]. These phonological rules are invoked to imitate the pronunciation of humans. Deriving these phonological rules, however, is laborsome, time-wasting and tedious. Besides, accumulating appropriate and complete rules is extremely difficult since the phonological characteristics are interactively affected by many linguistic features. Consequently, approaches using neural networks were investigated for automatically learning prosodic information [8]-[10]. However, problems involving large training data are associated with the unendurable long training time and, particularly, with the difficulty to converge.

This paper proposes a Chinese text-to-speech conversion system which focuses on the selection of synthesis unit and word prosodic pattern. An important characteristic of Mandarin Chinese is that it is a tonal language based on monosyllables. 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 perspective of Chinese phonology, the total number of phonologically allowed syllables in Mandarin speech is only around 1410. However, obstructed by the storage problem, a set of 408 syllables with high-level tone was generally used [4][5]. Such an approach might obtain a less satisfactory result of listening test because making substantial changes to the tonal manifestation of a syllable depends on the context. In order to process the coarticulation between concatenated syllables, synthesis units are extracted from a large continuous speech database. On the other hand, word is chosen as the unit for prosody modification because word is the basic rhythmical pronunciation unit. 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. Consequently, a word prosody database is established based on above linguistic features. Each word prosodic pattern contains the syllable duration, energy contour and pitch contour of the word. To represent the prosody con-

*The work was supported by National Science Council, Republic of China, under Grant NSC84-2622-E006-006

tours, the cubic spline curve fitting method [11] was used, which describes the contour by piecewise cubic polynomials.

2. System description

Fig. 1 presents the block diagram of this text-to-speech system, which is described in detail in the following.

- Text analysis: A preliminary text analysis is first used to identify the punctuation marks, numerals and Chinese characters. Meanwhile, some contextual features such as phonetic structure and syntactical structure are extracted. Also, a Chinese word dictionary having approximately 80,000 entries is available, in which the words are organized as a tree-like structure for improving the search speed. In the word segmentation process, a Chinese word segmentation module is invoked to segment character sequence into word sequence using the word dictionary. In the word formation process, monosyllable word not belonging to any word is combined with its preceding word to give the final word sequence.
- Phonetic Transcription: The phonetic transcription for each character is obtained by referring to a phonetic symbol-to-sound table.
- Word prosody selection: According to the information of the first two modules, a word prosodic pattern is chosen from the word prosody database based on the tone combination, part of speech and word position in a sentence.
- Synthesis unit selection: Synthesis units with varied prosodic and spectral characteristics from a large, single speaker speech database are selected and concatenated to generate synthesized speech. The selection of units is based on minimizing acoustic distortion between the selected units and desired units.
- Prosody modification module: By using the selected synthesis units and word prosodic patterns, the pitch-synchronous overlap-and-add (PSOLA) approach [6] is applied to adjust the prosody including the syllable duration, energy, pitch contour and pause duration.

3. Prosodic pattern selection

3.1 Word prosody database

In Chinese TTS system, some linguistic features are relevant to word-based prosodic information synthesis. They are: (1) tone combination, (2) word length, (3) part of speech of the word, and (4) word position in a sentence. These features are discussed in more detail in the following.

(1) Tone combination: Mandarin Chinese is a tonal language, including four lexical tones and one neutral tone; and each tone can be described in terms of the pitch contour. Also, the word is the basic comprehensive pronunciation unit. 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: In Mandarin speech, word is the basic pronunciation unit. There exist monosyllabic, disyllabic, and polysyllabic words. The coarticulation within a word is more obvious than that between two words. Therefore, this feature represented by length n for an n-syllable word is used to choose corresponding word patterns.

(3) Part of speech of the word: Part of speech (POS) is also an important linguistic feature to determine the word prosody. In this paper, the POS is divided into 21 categories. The distance between two categories is defined as the distance of their corresponding 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: In general, the pitch contour and energy contour in a sentence will follow an intonation pattern. For example, the pitch contour and the energy contour will decline in a declarative sentence. Therefore, the word position in a sentence will affect the word prosodic information. The word position in a sentence is defined as the word position divided by the number of words in the sentence. For example, the word position for the first word in a five-word sentence is 1/5.

Based on the above linguistic features, a word prosody database is constructed. Each word pattern in the word prosody database contains linguistic features, i.e., tone combination, word length, POS of the word and word position in a sentence, and prosodic features, i.e., pitch contour, energy contour and syllable duration.

To establish the word prosody database, a continuous speech database established by the Telecommunication Laboratories 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 then manually done. A total number of 38907 syllables and their phonetic labels were obtained. This yielded 9698 word patterns (including 2-, 3-, and 4-syllable words) after text analysis. In Table 1, the number of word patterns and the tone combinations in the database are shown for each n-syllable word, $2 \leq n \leq 4$. Note that the words with length above 4 are not in the table for their large number of tone combinations.

3.2 Pattern selection

In the word prosodic pattern selection process, the following information can be obtained for each word: tone combination of the word, word length, POS of the word and word position in a sentence. For each input word, the tone combination and word length

Table 1: Distribution of word prosody patterns in the speech database.

	Tone combinations	No. of word patterns
2-syllable word	20	5171
3-syllable word	80	2577
4-syllable word	320	1950

were first used to choose its corresponding word pattern candidates. Second, the POS and position of the input word were used to calculate the distance between the linguistic features of the input word and word patterns in the database using the POS distance table. The prosodic pattern with the minimum distance is chosen as the output.

$$j^* = \min_j \{d_C(C_I, C_j) + d_P(P_I, P_j)\}, \quad j = 1, \cdots, J \tag{1}$$

where J is the number of word pattern candidates corresponding to a given tone combination and word length. $d_C(C_I, C_j)$ represents the distance between the POSs of the input word C_I and the word pattern C_j in the database. $d_P(P_I, P_j)$ represents the absolute distance between word position of the input word P_I and word position of the word pattern P_j.

4. Synthesis unit selection

4.1 Synthesis unit inventory

To construct the synthesis unit inventory, the speech database described in Section 3.1 was used to obtain a set of 38907 syllable. The following two criteria are employed to filter out some of them.

- The syllables with duration less than 200ms were discarded because distinguishable distortion will be created in duration modification of short syllables.
- Some syllables were abundant in coarticulation, and one of them was selected as the synthesis unit. In Chinese, a syllable is composed of an optional initial/consonant followed by a final/vowel. Since syllable is chosen as the synthesis unit, the coarticulation between two concatenated syllables is affected by final followed by an initial or a final. In order to produce a natural realization of two concatenated syllables, both the prosodic and phonetic appropriateness of units should be considered. Table 2 and Table 3 display the groups of initials and finals according to the spectral similarity respectively. For those syllables followed by the same group of initials or finals, one of them was chosen as the synthesis unit which minimizes the distortion between its pitch contour and the average one.

On the other hand, it is difficult to construct a set of synthesis units comprising all possible combinations

Table 2: The four groups of the initials.

1	2	3	4
ㄇ (m)	ㄅ (b)	ㄐ (ji)	ㄓ (j)
ㄏ (h)	ㄆ (p)	ㄑ (chi)	ㄔ (ch)
ㄋ (n)	ㄈ (f)	ㄒ (shi)	ㄕ (sh)
ㄌ (l)	ㄉ (f)		ㄗ (tz)
ㄖ (r)	ㄊ (t)		ㄘ (ts)
	ㄍ (g)		ㄙ (s)
	ㄎ (k)		

Table 3: The five groups of the finals.

1	2	3	4	5
Null	ㄚ (a)	ㄢ (an)	ㄧ (yi)	ㄨ (wu)
	ㄛ (o)	ㄣ (en)	ㄞ (ai)	ㄠ (au)
	ㄜ (e)	ㄤ (ang)	ㄟ (ei)	ㄡ (ou)
	ㄦ (er)	ㄥ (eng)	ㄩ (yu)	
			ㄝ (eh)	

of coarticulations. A strategy is proposed to find the unit minimizing the spectral distortion between possible concatenated units. The cepstral parameters have been adopted to estimate spectral distortion [2][3]. In this paper, the LSP frequencies are employed because they are similar to formant frequencies and have small spectral resolution variation [12]. The evidence can be revealed from Fig. 2 which illustrates the LPC coefficients, cepstral coefficients and LSP frequencies of the utterance "Taiwan" respectively. For each synthesis unit, the inventory contains the following information.

- the waveform,
- the phonemic symbols of the preceding, the following and the current syllables,
- and 10 LSP frequencies of the first and the last frames.

4.2 Unit selection

In the unit selection phase, given an input syllable S with preceding syllable A and succeeding syllable B, the target synthesis unit is selected as follows:

- Search the unit inventory to find a syllable S with A the preceding syllable and B the succeeding syllable. If there is an exact match, then the corresponding stored waveform is selected as the target synthesis unit.
- Otherwise calculate the inter-syllable spectral distortion by

$$d_{ASB} = \sum_{m=1}^{P} [(w_{lm}^A - w_{fm}^S)^2 + (w_{lm}^S - w_{fm}^B)^2], \tag{2}$$

where P is the LSP order; w_{fm}^T and w_{lm}^T denote the mth LSP frequencies of the first frame and

Table 4: Results for the intelligibility test.

	Amount	Intelligibility
2-syllable word	100	93.8%
3-syllable word	100	96.6%
4-syllable word	100	97.4%
Sentence	50	97.9%
Short text	5	98.2%
Average		96.8%

the last frame of syllable T respectively. By (2), the selected synthesis unit is the one with minimum distortion.

5. Performance evaluation

This text-to-speech system was implemented on a PC/AT 586 computer. Some preliminary performance evaluations were made on this system. Ten subjects were requested to subjectively evaluate this TTS system using the following criteria.

1. Intelligibility: In this test, the subjects were requested to listen to the synthesized speech without prior knowledge of the content. Next, the subjects wrote down what they heard. By comparing the results with original text, the correct rate was obtained.

2. Naturalness: At first, the subjects were asked to listen to two types of speech, respectively pronounced by a person and by a TTS system without prosody modification. The synthesized speech with prosody modification using the proposed TTS system was then evaluated. For the synthesized speech, the subjects gave mean opinion scores (MOS) on a scale of 1 to 5, i.e., 5 for excellent level, 4 for good level, 3 for fair level, 2 for poor level, and 1 for unsatisfactory level.

The evaluation for the intelligibility test is shown in Table 4. The average correct rate is 96.8%. As indicated in this table, a word with longer length is more intelligible since it includes more semantic information. On the other hand, some words with fricative initials are inherently confusable in pronunciation, such as the initials 'j', 'ch' and 'sh' vs. the initials 'tz', 'ts', and 's' respectively. This factor largely increases the error rates. Table 5 lists the MOS's for words or sentences with different lengths. As indicated in this table, the average MOS is 3.7 for naturalness. As opposite to the intelligibility test, the results indicate that a shorter token length obtains a higher MOS since less linguistic information is needed. Furthermore, the MOS for short text is lower than the average MOS. The reason is the lack of syntactic and semantic information for providing more prosodic information in this system.

6. Conclusion

Table 5: Results for the naturalness test.

	Amount	Naturalness(MOS)
2-syllable word	100	3.8
3-syllable word	100	3.7
4-syllable word	100	3.7
Sentence	50	3.6
Short text	5	3.5
Average		3.7

This paper presents approaches for the selection of synthesis unit and word prosodic pattern in a Chinese Text-to-Speech system using a large speech database. A set of multiple candidates is adopted as the basic synthesis units for each syllable. For each word pattern, prosodic patterns, including pitch contour, energy contour, and word duration, was stored in a word prosody database generated from a large speech database. The cubic spline curve fitting method was adopted to approximate the dynamics of prosody contour. On the other hand, the LSP frequencies were applied to estimate the spectral distortion in the unit selection. Evaluation by subjective experiments confirmed the satisfactory performance of this approach.

Acknowledgements

The authors would like to thank the Telecommunication Laboratories for providing of the speech database.

References

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

[2] W. J. Wang, *et al*, "Tree-based unit selection for speech synthesis," ICASSP, pp.II-191-II-194, 1993.

[3] N. Iwahashi and Y. Sagisaka, "Speech segment network approach for optimization of synthesis unit set," Computer Speech and Language, pp.335-352, 1995.

[4] 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, Vol.37, No.9, pp.1309-1319, Sept. 1989.

[5] L. S. Lee, C. Y. Tseng, and C. J. Hsieh, "Improved tone concatenation rules in a formant-based Chinese text-to-speech system," IEEE Trans. on Speech And Audio Processing. Vol. 1, No. 3, July, 1993.

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

[7] D. Bigorgne, *et al*, "Multilingual PSOLA text-to-speech system," ICASSP, pp.II-187-II-190, 1993

[8] M. S. Scordilis and J. N. Gowdy, "Neural network based generation of fundamental frequency contours," ICASSP, pp.219-222, 1989.

[9] M. G. Rahim and C. C. Goodyear, "Articulatory synthesis with the aid of a neural net, " ICASSP, pp.227-230, 1989.

[10] S. H. Chen, S. H. Hwang, and C. Y. Tsai, "A first study on neural net based generation of prosodic and spectral information for mandarin text-to-speech," ICASSP, pp.45-48, 1992.

[11] Gerald, *Applied numerical analysis* Addison-Wesley Publishing Company, Inc, pp.290-294, 1973.

[12] S. Furui, *Digital speech processing, synthesis, and recognition* Marcel Dekker, Inc, p.134, 1989.

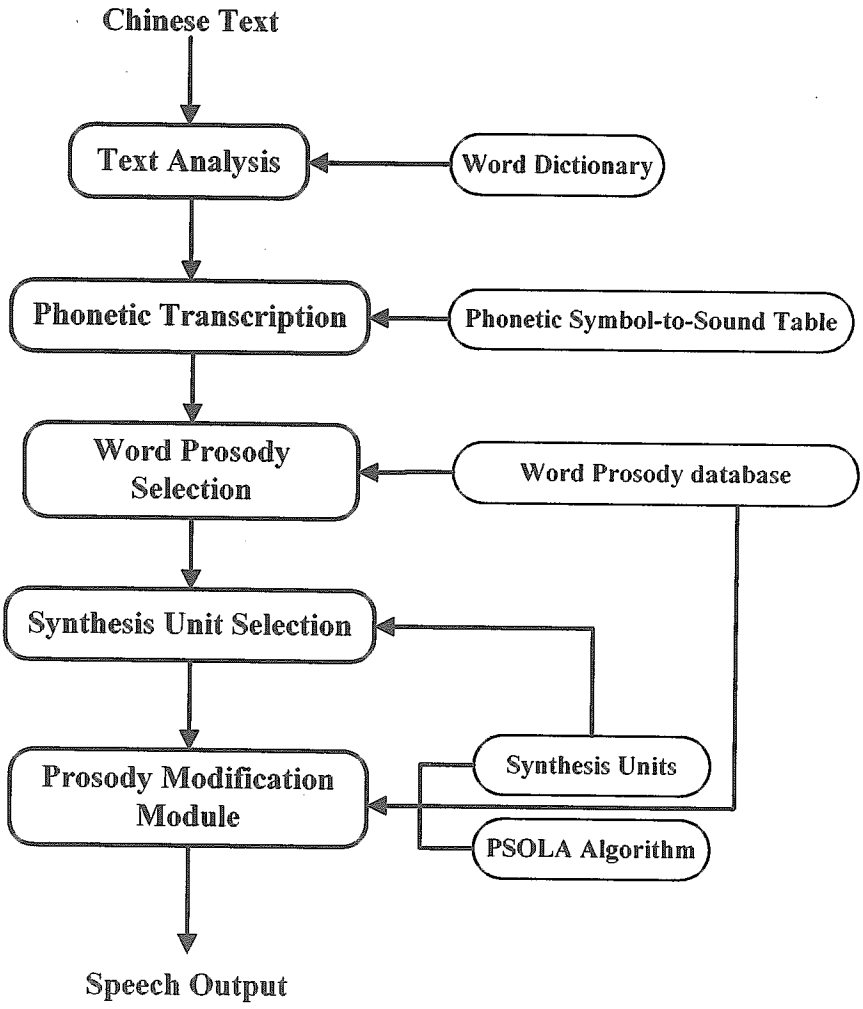

Fig. 1. The block diagram of the TTS system.

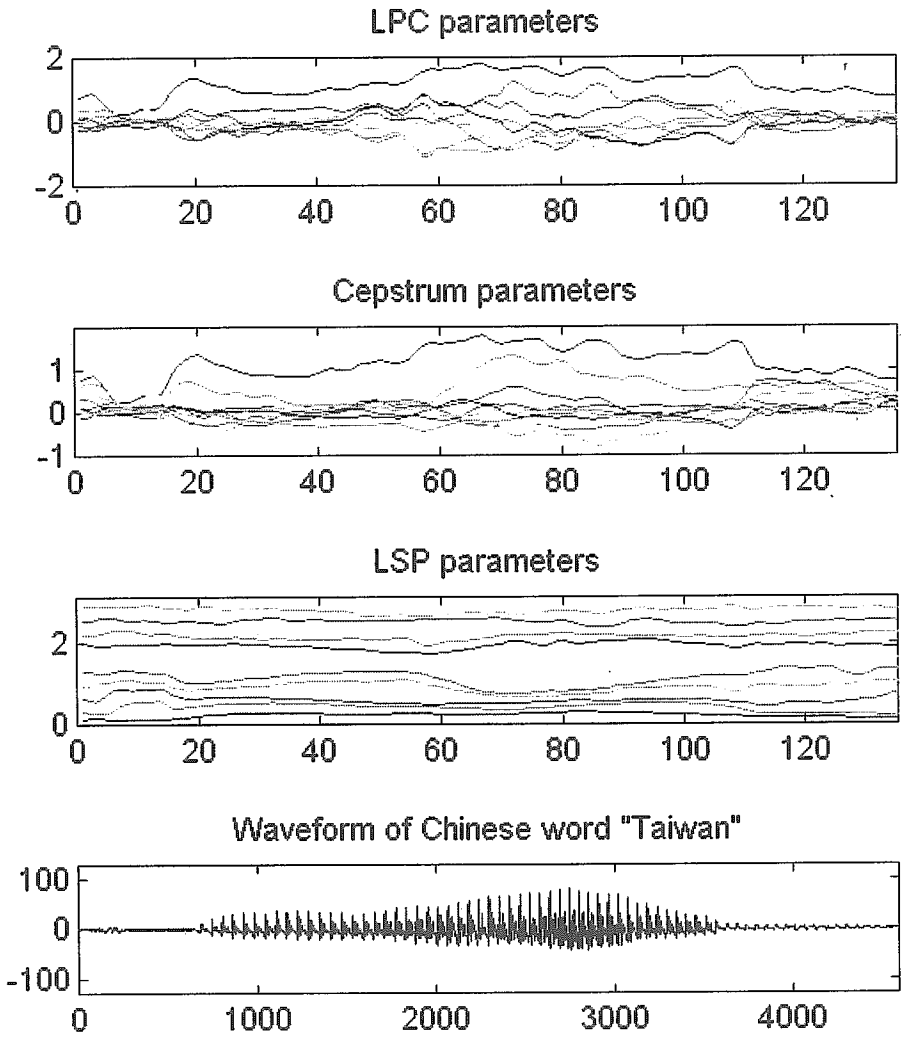

Fig. 2. The spectral contours of the utterance "Taiwan" of the LPC , cepstrum and LSP parameters.