ELSEVIER

Speech Communication 35 (2001) 219–237

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic generation of synthesis units and prosodic information for Chinese concatenative synthesis

Chung-Hsien Wu *, Jau-Hung Chen

Department of Computer Science and Information Engineering, National Cheng Kung University, 1 Ta-Hsueh Road, Tainan, Taiwan, ROC

Received 17 August 1999; accepted 25 July 2000

Abstract

In this paper, some approaches to the generation of synthesis units and prosodic information are proposed for Mandarin Chinese text-to-speech (TTS) conversion. The monosyllables are adopted as the basic synthesis units. A set of synthesis units is selected from a large continuous speech database based on two cost functions, which minimize the inter- and intra-syllable distortion. The speech database is also 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 phrase. 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 showed that the synthesized prosodic features matched quite well with their original counterparts. Evaluation by subjective experiments also confirmed the satisfactory performance of these approaches.

Keywords: Chinese text-to-speech conversion; Synthesis units; Prosodic information; Concatenative synthesis; Pitch contour; Syllable duration

1. Introduction

In past years, many studies have focused on TTS systems for different languages (Klatt, 1987; Bigorgne et al., 1993; Kawai et al., 1995). Also, TTS systems and synthesis technology for the Chinese language have been developed in the last two decades (Lee et al., 1989, 1993; Chan and Chan 1992; Chen et al., 1998; Shih and Sproat, 1996; Chou and Tseng, 1998). A detailed overview of TTS systems for English and Chinese were introduced by Klatt (1987) and Shih and Sproat (1996), respectively. Potential applications include aids for the handicapped, teaching aids, speech-to-speech translation, and any applications for text reading, such as email reader, news reader, and so on.

General speaking, a TTS system could be logically composed of three main parts: text/linguistic analysis, prosodic information generation, and speech synthesis. Text analysis is first invoked to analyze the input text. It carries out two main tasks: (1) transcription from letter to phoneme is executed, which forms the basis for the following acoustic unit selection and (2) morphologic and syntactic analyses are performed to extract contextual information and output linguistic features.

* Corresponding author. Tel.: +86-2757575; fax: +86-2747076.

0167-6393/01/$ - see front matter
PII: S0167-6393(00)00075-3

Generally, text analysis is a language-dependent component in a TTS system. The prosodic information generation employs the linguistic features to generate prosodic features including pitch contour, energy contour, and duration. The prosodic features are controlled to represent the prosodic/suprasegmental information of spontaneous speech, such as time-varying curve of fundamental frequency (F_0), accent and stress, pausing rhythm, and intonation. This information also depends on the attitudinal and emotional connotations, speaking styles, speaker characteristics, and referential relations. Finally, speech synthesis is performed to modify the prosodic parameters of the synthesis units and generate intelligible and natural speech based on the above features. There are three modern approaches to speech synthesis: articulatory synthesis, formant synthesis, and concatenative synthesis (Shih and Sproat, 1996). The concatenative synthesis is the simplest and effective approach, which uses real recorded speech as the synthesis units and concatenates them back together during synthesis. This approach is adopted by most of the TTS systems today. Besides, the pitch-synchronous overlap-and-add (PSOLA) approach (Charpentier and Stella, 1986) or other overlap-and-add approaches (Quatieri and McAulay, 1992; George and Smith, 1997) are used to adjust the prosodic features in time/frequency domain.

In concatenative speech synthesis, unit selection plays a prominent role of synthesizing intelligible, natural and high-quality speech. In past years, many kinds of synthesis units have been proposed (Klatt, 1987). The phonemes have been adopted as the basic synthesis units. Such units take advantage of small storage. However, it needs to improve the accuracy of intra-syllable coarticulation and the spectral discontinuity between adjacent units. Consequently, longer synthesis units, such as diphone, demi-syllable, syllable, triphone and polyphone, are adopted to reduce the effect of spectral distortion (Klatt, 1987; Bigorgne et al., 1993). Recently, the approaches to unit selection from a large speech database or using non-uniform units have been appreciated and proved to obtain natural and high-quality speech (Chou and Tseng, 1998). This approach defined a cost function to select an appropriate sequence of synthesis segment.

On the other hand, rule-based approach has been used for prosody modification (Klatt, 1987; Lee et al., 1989, 1993; Chan and Chan, 1992). These phonological rules are invoked to imitate the pronunciation of humans. The derivation of phonological rules, however, is labor intensive 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 (Chen et al., 1998; Scordilis and Gowdy, 1989). However, the network may 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 Mandarin Chinese text-to-speech (TTS) conversion system, which focuses on the generation of synthesis units and prosodic information. 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 mid-falling–rising tone (Tone 3), the high-falling tone (Tone 4), and the neutral tone (Tone 5). From the viewpoint of Mandarin Chinese phonology, the total number of phonologically allowed syllables in Mandarin speech are only about 1300. Therefore, a syllable is a linguistically appealing synthesis unit in a Mandarin Chinese TTS system. However, due to the storage problem, a set of 408 syllables with the high-level tone has generally been used (Lee et al., 1989). Such an approach might obtain less satisfactory results for the intelligibility test because substantial changes in the tonal manifestations of a syllable depending on the context. 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. Two types of distortion measure are employed in the determination of synthesis units. One is the intra-syllable distortion (syllable cost), which represents the distances

between speech units with the same syllable. The other is the inter-syllable distortion (concatenation cost), which is the measure of the spectral continuity between two adjacent syllables. They will be explained in detail in the next section

On the other hand, 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 Mandarin Chinese word is generally affected by the tone combination, word length, POS of the word, and word position in a phrase. In this paper, 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. The pitch contour in the word-prosody template records the well-known tone sandhi for the syllables in 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 PSOLA method is employed for the modification of prosodic information (Charpentier et al., 1986).

The rest of this paper is organized as follows. The system is described in Section 2. The proposed approach to synthesis unit selection is introduced in Section 3. In Section 4, the template-based prosody generation is presented. Experimental results are provided in Section 5. Concluding remarks are finally made in Section 6.

2. System description

The block diagram of the TTS system is shown in Fig. 1. It is divided into two main parts: training part and synthesis part. They are described in detail in the following:

Training part. A large continuous speech database, designed and provided by Telecommunication Laboratories, MOTC, Taiwan, containing 655 reading texts is used to generate the synthesis unit inventory and the prosodic information inventory. Five procedures are proposed to select a

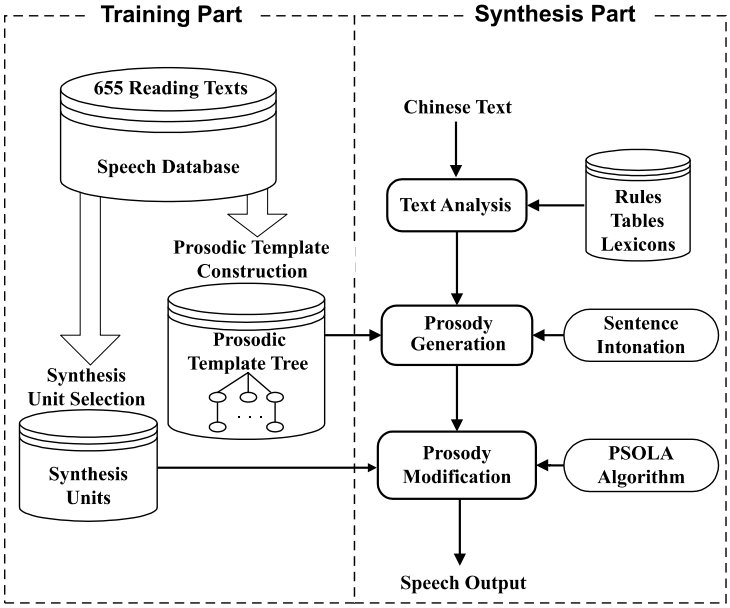

Fig. 1. Block diagram of the TTS system.

set of synthesis units from the speech database: pitch-period detection and smoothing, speech segment filtering, spectral feature extraction, unit selection, and manual examination. The procedure of unit selection is based on two cost functions, which minimize the inter- and intra-syllable distortion. They will be explained in detail in the next section.

The structure of the prosodic information inventory is a word-prosody template tree, which is constructed according to the linguistic features: tone combination, word length, POS of the word, and word position in a phrase. This template tree stores the prosodic features including pitch contour, average energy, and syllable duration of a word for possible combinations of linguistic features.

Synthesis part. For an input text, the following modules are invoked to output the synthesized speech:

1. *Text analysis*. The input text is an arbitrary Big5 code string (for Mandarin Chinese). Text analysis is first performed to identify punctuation, Arabic numerals, and Mandarin Chinese characters. The syntactic structure is extracted by a simple partial parser and used to deal with homograph disambiguation. Also, a Mandarin Chinese word dictionary of about 80,000 entries is used for word identification. Dictionary look-up of word pronunciation and rules for homonym disambiguation are performed. The phonetic transcription for each character is obtained by referring to a phonetic table.
2. *Prosody generation*. A sentence intonation module is first used to compute the target pitch periods of a word in the word sequence. The linguistic features of each word and the target pitch periods are then fed to a template selection module. Based on the target pitch periods, this module uses a cost function to estimate the distance of linguistic features between the input word and the one in the template tree and output the prosodic features from the word template tree.
3. *Prosody modification*. Using the outputs from the prosodic information generator, prosody modification based on the PSOLA approach is carried out to produce synthesized speech. It adjusts the word prosody including the syllable duration, energy contour and pitch contour.

3. Synthesis unit selection

Fig. 2 shows the block diagram of the synthesis unit selection. Each block in this diagram is described in detail as follows.

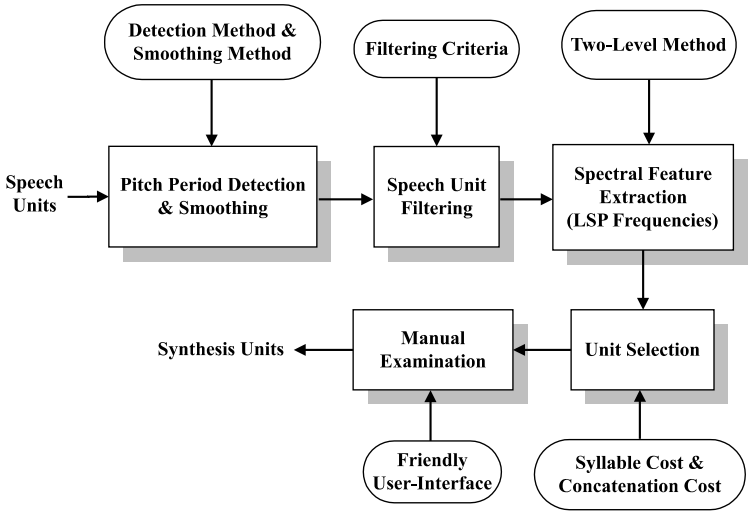

Fig.2. Block diagram of the synthesis unit selection.

3.1. Pitch-period detection and smoothing

For each speech unit in the speech database, pitch-mark labeling is automatically estimated by the autocorrelation method (Rabiner and Schafer, 1978) and a modified C/V segmentation algorithm (Wang et al., 1991). Besides, quantitative description of the pitch contours is expressed by orthonormal expansion using discrete Legendre polynomials (Chen and Wang, 1990). That is, a pitch contour can be represented by a four-dimensional vector (a_0, a_1, a_2, a_3) and it is referred to as *pitch vector* in this paper.

It is inevitable that a few errors of pitch periods remain in the pitch contours. Some of these errors are due to the fact that the second peak is greater than the first peak in a pitch period. This type of errors can be effectively removed by the reconstruction of pitch periods from pitch vector. The other errors are mostly caused by the disturbance at the beginning or end of voiced parts, which make significant pitch jumps. For this type of errors, a simple smoothing method (Rabiner and Schafer, 1978) is used to eliminate the discontinuity at both the beginning and end in a pitch contour.

Fig. 3 illustrates an example of pitch contours before and after smoothing. It can be seen that the original pitch contour has conspicuous errors at the end part. A smoothed pitch contour is obtained after the reconstruction of pitch contour by discrete Legendre polynomials. As seen in this figure, this method is able to eliminate the errors due to gross measurement and adequately smooth the contour.

3.2. Speech unit filtering

For some reasons, such as units with few pitch marks, some speech units are not qualified to be the synthesis units. Therefore, it is necessary and important to filter them out before further processing. Four criteria are adopted in this module and described below.

Syllable duration. Generally, short units are not pronounced well or completely. They are not intelligible and will generate distinguishable distortion in duration modification. In our system, the speech units with duration less than 170 ms were discarded.

Total number of pitch marks. Some syllables with few pitch marks result from the unstable or short voiced parts. They are not suitable for duration or pitch modification. A threshold of eight pitch marks was set in this module.

Inter-syllable coarticulations. Inter-syllable coarticulations capture the transitions between syllables and are helpful for generating natural speech. However, synthesis units with strong inter-syllable coarticulations are not appreciated for the following two reasons:

1. Additional articulations at the syllable boundaries are notably perceived when read individually.

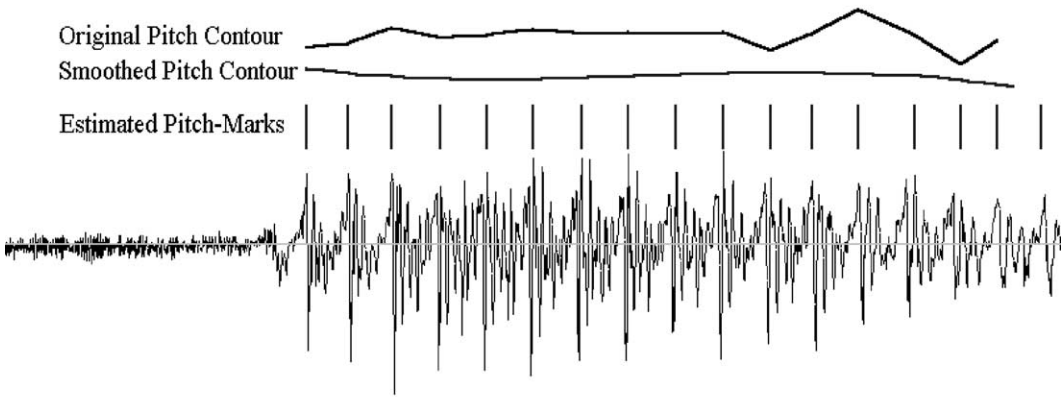

Fig. 3. An example of pitch contour smoothing using discrete Legendre polynomials.

2. They are not harmonically concatenated with other units having discrepant inter-syllable coarticulations.

It is known that two concatenative voiced speech units have strong inter-syllable coarticulations. Therefore, three types of inter-syllable concatenation are defined and displayed in Table 1 by investigating the inherent phonetic properties of phonemes. The term "loose concatenation" indicates few effects of inter-syllable coarticulations. Therefore, the syllables of this type are first chosen as the candidates of synthesis units. On the other hand, tight and overlapped concatenations have medium and strong effects of inter-syllable coarticulation, respectively. Consequently, they are not in the top-priority list. An example of the three concatenation types is shown in Fig. 4. The vertical dash-lines indicate syllable boundaries. It can be seen that the first two syllables with overlapped concatenation are blurred together and hence very difficult to obtain a precise syllable boundary. The following two syllables have a short silence in between and they are labeled as loose concatenation.

Table 1
Three types of inter-syllable concatenation

The last phoneme of the preceding syllable	The first phoneme of the current syllable	Types of inter-syllable concatenation
j(ㄓ), ch(ㄔ), sh(ㄕ), r(ㄖ), tz(ㄗ), ts(ㄘ), s(ㄙ), a(ㄚ), o(ㄛ), e(ㄜ), eh(ㄝ), ai(ㄞ), ei(ㄟ), au(ㄠ), ou(ㄡ), an(ㄢ), en(ㄣ), ang(ㄤ), eng(ㄥ), er(ㄦ), yi(一), wu(ㄨ), yu(ㄩ)	Unvoiced initials: b(ㄅ), p(ㄆ), f(ㄈ), d(ㄉ), t(ㄊ), g(ㄍ), k(ㄎ), h(ㄏ), ji(ㄐ), chi(ㄑ), shi(ㄒ), j(ㄓ), ch(ㄔ), sh(ㄕ), tz(ㄗ), ts(ㄘ), s(ㄙ)	Loose concatenation
	Voiced initials: m(ㄇ), n(ㄋ), l(ㄌ), r(ㄖ)	Tight concatenation
	Voiced finals: a(ㄚ), o(ㄛ), e(ㄜ), eh(ㄝ), ai(ㄞ), ei(ㄟ), au(ㄠ), ou(ㄡ), an(ㄢ), en(ㄣ), ang(ㄤ), eng(ㄥ), er(ㄦ), yi(一), wu(ㄨ), yu(ㄩ)	Overlapped concatenation

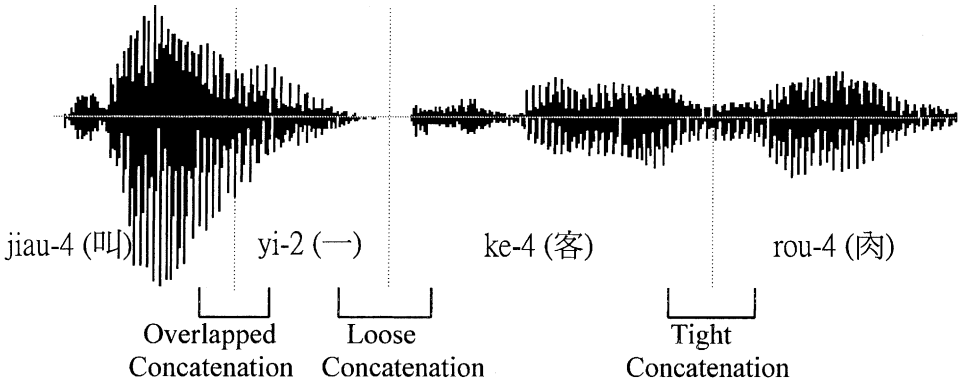

Fig. 4. Example of the three concatenation types.

The following steps are used to filter out undesired units:

Step 1. Initialization:
Let A_i denote the set of speech units for syllable i in which the unsuitable ones have been filtered out by the above two criteria. Also, let $B_i = \emptyset$.

Step 2. Add units to B_i:
Select all the units with loose concatenation type from A_i and add them to B_i. If $|B_i|$ is greater than a threshold N, then go to Step 3. Otherwise, this step is continued for tight concatenation and then overlapped concatenation. In this paper, N is set to 6.

Step 3. Quit.

Ranges of pitch period and syllable intensity. Syllables with extremely large/small pitch periods are unsuitable for pitch modification. According to our investigation, the desired average pitch periods of syllables are between 5 and 11 ms for the male speaker. Also, the desired syllable intensities are between half and twice of the average intensity for each syllable. The filtering process is as follows:

Step 1. Initialization:
Set $C_i = B_i$, where B_i represents the unit set obtained from the above filtering process.

Step 2. Remove units from C_i:
If one of the following conditions is true, then go to Step 3.
1. $|C_i| < N$.
2. The pitch periods and syllable intensities of the speech units in C_i meet the above constraints.

Otherwise, remove a speech unit from C_i by the following priorities in turn:
(a) Smallest pitch period.
(b) Largest pitch period.
(c) Smallest syllable intensity.
(d) Largest syllable intensity.

Continue this step.

Step 3. Quit.

Finally, the set C_i contains the desired speech units.

3.3. Spectral feature extraction

In a concatenative TTS system, the spectral contours of two adjacent syllables should be as smooth as possible to generate natural speech. In order to select generic synthesis units, it is necessary to minimize the inter-syllable spectral distortion. In our system, the line spectrum pair (LSP) frequencies are adopted as the spectral parameters because they are similar to formant frequencies and have small spectral resolution variation. A detailed introduction and estimation of the LSP frequencies can be found in (Wu and Chen, 1997). For each speech unit, 10 frames are extracted in which each frame contains 10 LSP frequencies using a window of 25 ms. They are then used to estimate the spectral distortion in unit selection.

3.4. Automatic unit selection

Two types of distortion measure are employed in the determination of synthesis units: syllable cost (intra-syllable distortion) and concatenation cost (inter-syllable distortion). They are described as follows.

Syllable cost. Four kinds of features are used to estimate the syllable cost: the LSP frequencies, pitch vector, average intensity, and duration. The syllable cost is a weighted sum of the distance between the feature vectors of the input units and their mean vectors (center). For an input unit s_i of syllable j, the syllable cost function is represented as

$$\mathrm{SC}_j(s_i) = \omega^{\mathrm{F}} \cdot D_{\mathrm{LSP}}(F_i, \bar{F}_j) + \omega^{\mathrm{P}} \cdot D_{\mathrm{P}}(P_i, \bar{P}_j) + \omega^{\mathrm{I}} \cdot D_{\mathrm{I}}(I_i, \bar{I}_j) + \omega^{\mathrm{D}} \cdot D_{\mathrm{D}}(D_i, \bar{D}_j), \quad (1)$$

where the notations are described as follows:

F_i is the LSP frequency of s_i, P_i the pitch vector of s_i, I_i the average intensity of s_i, D_i the duration of s_i, $\bar{F}_j$, $\bar{P}_j$, $\bar{I}_j$ and $\bar{D}_j$ are the mean vectors of their corresponding features of syllable j. $D_{\mathrm{LSP}}(F_i, \bar{F}_j)$ is the distance measure for the LSP frequencies, which is defined as

$$D_{\mathrm{LSP}}(F_i, \bar{F}_j) = \frac{1}{FN} \left\{ \sum_{k=1}^{FN} \sum_{m=1}^{M} \left(\frac{F_{ikm} - \bar{F}_{jkm}}{\sigma_{km}^{F_j}} \right)^2 \right\}^{1/2}, \quad (2)$$

where FN is the total frame number, M the order of LSP frequencies, and σ^{F_j} is the standard deviation

of F_j. $D_{\mathrm{P}}(P_i,\bar{P}_j)$ is the distance measure for the pitch vector, which is defined as

$$D_{\mathrm{P}}(P_i,\bar{P}_j)=\left\{\sum_{m=0}^{3}\left(\frac{a_{im}-\bar{a}_{jm}}{\sigma_m^{P_j}}\right)^2\right\}^{1/2}, \tag{3}$$

where σ^{P_j} is the standard deviation of P_j. $D_{\mathrm{I}}(I_i,\bar{I}_j)$ is the distance measure for intensity, which is defined as the sum of the initial part and the final part,

$$D_{\mathrm{I}}(I_i,\bar{I}_j)=\left|\frac{I_{i0}-\bar{I}_{j0}}{\sigma_0^{I_j}}\right|+\left|\frac{I_{i1}-\bar{I}_{j1}}{\sigma_1^{I_j}}\right|, \tag{4}$$

where σ^{I_j} is the standard deviation of I_j. $D_{\mathrm{D}}(D_i,\bar{D}_j)$ is the distance measure for duration, which is defined as the sum of the initial part and the final part,

$$D_{\mathrm{D}}(D_i,\bar{D}_j)=\left|\frac{D_{i0}-\bar{D}_{j0}}{\sigma_0^{D_j}}\right|+\left|\frac{D_{i1}-\bar{D}_{j1}}{\sigma_1^{D_j}}\right|, \tag{5}$$

where σ^{D_j} is the standard deviation of D_j. $\omega^{\mathrm{F}},\omega^{\mathrm{P}},\omega^{\mathrm{I}}$ and ω^{D} are the weights of their corresponding features.

Concatenation cost. In this paper, the LSP frequencies and concatenation types are used to calculate the spectral distortion at syllable boundaries. For a syllable, the concatenation cost includes two spectral distortions: left concatenation distortion and right concatenation distortion. Left concatenation distortion is the distortion between the last frame of the preceding syllable and the first frame of the current syllable, while right concatenation distortion is the distortion between the last frame of the current syllable and the first frame of the following syllable. For an input unit s_i belonging to syllable j, the concatenation cost function is represented as

$$\mathrm{CC}_j(s_i)=\frac{1}{2(N-1)}\sum_{k=1,k\neq i}^{N}\left[\frac{D_{\mathrm{l}}(F_i,\bar{F}_k)}{\omega_{\mathrm{l}}(s_i)}+\frac{D_{\mathrm{r}}(F_i,\bar{F}_k)}{\omega_{\mathrm{r}}(s_i)}\right], \tag{6}$$

where N is the total number of different syllables, $D_{\mathrm{l}}(\cdot)$ and $D_{\mathrm{r}}(\cdot)$ denote the Euclidean distance measure for spectral distortions of the left and right concatenation, respectively. It is noted that $D_{\mathrm{l}}(\cdot)$ and $D_{\mathrm{r}}(\cdot)$ only calculate the distances of the LSP frequencies in desired frames. $\omega_{\mathrm{l}}(s_i)$ and $\omega_{\mathrm{r}}(s_i)$ are two coarticulation weights with respect to concatenation types between syllable s_i and its left and right speech units. As mentioned above, the fewer effects of inter-syllable coarticulation a speech unit is the more possibility it is qualified as a synthesis unit. Therefore, the coarticulation weight is chosen as

$$\text{Coarticulation weight}=\begin{cases}\omega_{\mathrm{l}} & \text{for loose concatenation,}\\ \omega_{\mathrm{t}} & \text{for tight concatenation,}\\ \omega_{\mathrm{o}} & \text{for overlapped concatenation,}\end{cases} \tag{7}$$

where $\omega_{\mathrm{l}}>\omega_{\mathrm{t}}>\omega_{\mathrm{o}}$.

Finally, a speech unit is selected as the synthesis unit for syllable j minimizing the following total cost, which is a weighted sum of the syllable cost and the concatenation cost,

$$\mathrm{TC}_j(s_i)=\omega^{\mathrm{SC}}\cdot\mathrm{SC}_j(s_i)+(1-\omega^{\mathrm{SC}})\cdot\mathrm{CC}_j(s_i). \tag{8}$$

The above procedures are continued until all the synthesis units are selected.

3.5. Manual examination

Although a computer can select a set of synthesis units automatically, it is possible that some poor units are selected from speech units with few candidate samples. To obtain a set of synthesis units with better speech quality, each selected synthesis unit was inspected subjectively by an experienced person. Poor units were manually replaced by better ones.

For each synthesis unit, the inventory contains the following information:

- The waveform and its length.
- Average energies of the initial and the final parts.
- Pitch marks, total number of pitch marks, and average pitch period.
- Beginning position of the final part.
- Group number of the initial part.
- Group number of the final part.

The initial parts and final parts are clustered into some groups based on their phonetic characteris-

Table 2
Eight groups of the initial

Group	1	2	3	4	5	6	7	8
Initial	Null	m (ㄇ) n (ㄋ) l (ㄌ) r (ㄖ)	b (ㄅ) d (ㄉ) g (ㄍ)	f (ㄈ) h (ㄏ)	p (ㄆ) t (ㄊ) k (ㄎ)	ji (ㄐ) chi (ㄑ) shi (ㄒ)	j (ㄓ) ch (ㄔ) sh (ㄕ)	tz (ㄗ) ts (ㄘ) s (ㄙ)

Table 3
Ten groups of the final

Group	1	2	3	4	5
Final	a (ㄚ) ai(ㄞ) au(ㄠ)	an(ㄢ) en(ㄣ) ang(ㄤ) eng(ㄥ)	o (ㄛ) ou(ㄡ) wu(ㄨ)	yi(ㄧ) yu(ㄩ)	e(ㄜ) eh(ㄝ) ei(ㄟ) er(ㄦ)
Group	6	7	8	9	10
Final	ia (ㄧㄚ) iai(ㄧㄞ) iau(ㄧㄠ) ieh(ㄧㄝ) yue(ㄩㄝ) io (ㄧㄛ) iou(ㄧㄡ)	ian(ㄧㄢ) ien(ㄧㄣ) iang(ㄧㄤ) ieng(ㄧㄥ) uan(ㄨㄢ) yuen(ㄩㄣ) yung(ㄩㄥ)	ng1(ㄫ)	ua (ㄨㄚ) uai(ㄨㄞ) uo(ㄨㄛ) uo(ㄨㄛ)	uan(ㄨㄢ) uen(ㄨㄣ) uang(ㄨㄤ) ung(ㄨㄥ)

tics and the corresponding group numbers are listed in Tables 2 and 3, respectively. All the above information is stored for prosody modification.

4. Template-based prosody generation

In the Mandarin Chinese TTS system, some linguistic features are relevant to the information of word prosody. They are tone combination, word length, POS of the word, and word position in a phrase. 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 result in different F_0 contours, which might vary in shapes, slopes and means. The above results convince us that tone combination is a significant linguistic feature of estimating prosodic information. 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 $4^{n-1} \times 5$ tone combinations for an n-syllable word.
2. *Word length*. In Mandarin speech, the word is the basic pronunciation unit, i.e., the word is the basic unit of a phrase. Some of the prosodic characteristics of the phrase, such as final lengthening effect, are also presented in the word. Therefore, for an n-syllable word, the word length, n, is used to obtain the corresponding prosodic templates. The frequently used words are with the length of 1–4, that is, monosyllable, 2-syllable, 3-syllable and 4-syllable words.
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. They are listed in Table 4.

Table 4
The 18 types of POS

Non-predicative adjective (A)	Negation (DC)	Preposition (P)
General adverb (ADV)	Measure (M)	Determiner (DT)
Aspectual adverb (ASP)	Interjection (I)	General Noun (N)
Coordinate conjunction (CA)	Localizer (LOC)	Pronoun (NH)
Conjunction at the beginning of a sentence (CB)	Conjunction at the end of a sentence (CE)	Transitive/Intransitive verb (VTI)
Intransitive verb (VI)	Transitive verb (VT)	Other POSs (OP)

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. This implies that the word position in a phrase will affect the prosodic information.

In this paper, a word-prosody template tree is constructed based on the above linguistic features. The linguistic features of a word, i.e., tone combination, word length, POS of the word and word position in a phrase, are associated with a set of prosodic patterns, i.e., syllable duration, energy contour and pitch contour. To establish the word-prosody template tree, the same speech database used in previous section was adopted. 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 5.

4.1. Structure of the word-prosody template tree

The structure of the word-prosody template tree is shown in Fig. 5. 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 with respect to this syllable.
- The group numbers of the initial and the final parts.
- The codes of the initial and the final parts.
- Big5 codes.

It is noted that the above scheme does not take into account the linguistic feature "word position in a phrase". The main reason is that it is difficult to predict the phrase boundaries in the text analysis. However, the prosodic effects of word position in a phrase have relation with the trajectory of F_0 contours. A general result is that 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.

On the other hand, some of the pitch contours in the speech database strongly depend on the context or considerably deviate from their representatives. The latter could be resulted from either wrong F_0 estimation or something in relation to the speaker's speaking habit. To avoid comprising the word patterns with irregular or erroneous pitch

Table 5
Distribution of word-prosody patterns in the speech database

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

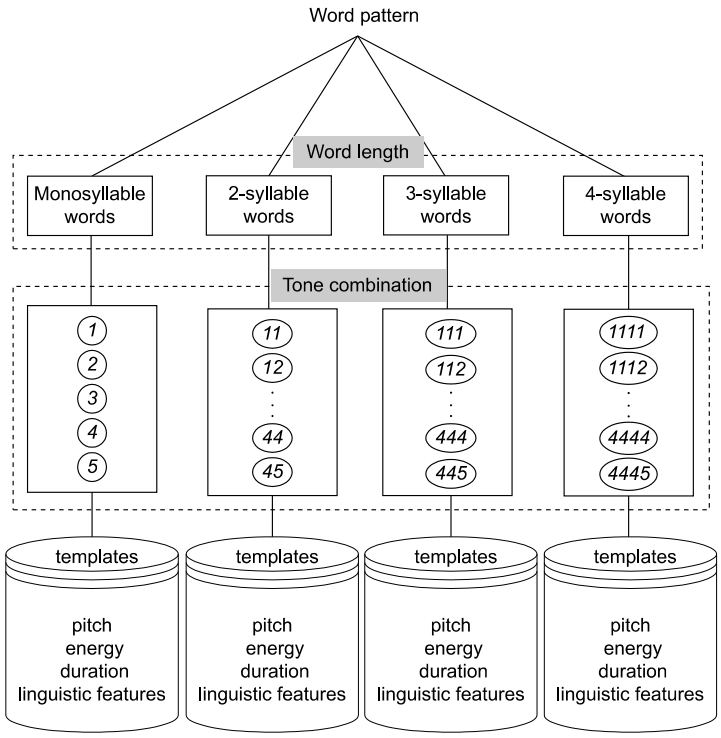

Fig. 5. Structure of the word-prosody template tree.

contours, the pitch vectors represented by (a_0, a_1, a_2, a_3) are used to eliminate this problem. The component a_0 represents average pitch period and the other three components are related to the shape of the pitch contour.

4.2. Generation of word-prosody templates

The generation of the word-prosody templates is shown in Fig. 6. 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 phrase. 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:

4.2.1. 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. In our system, it has been implemented, with respect

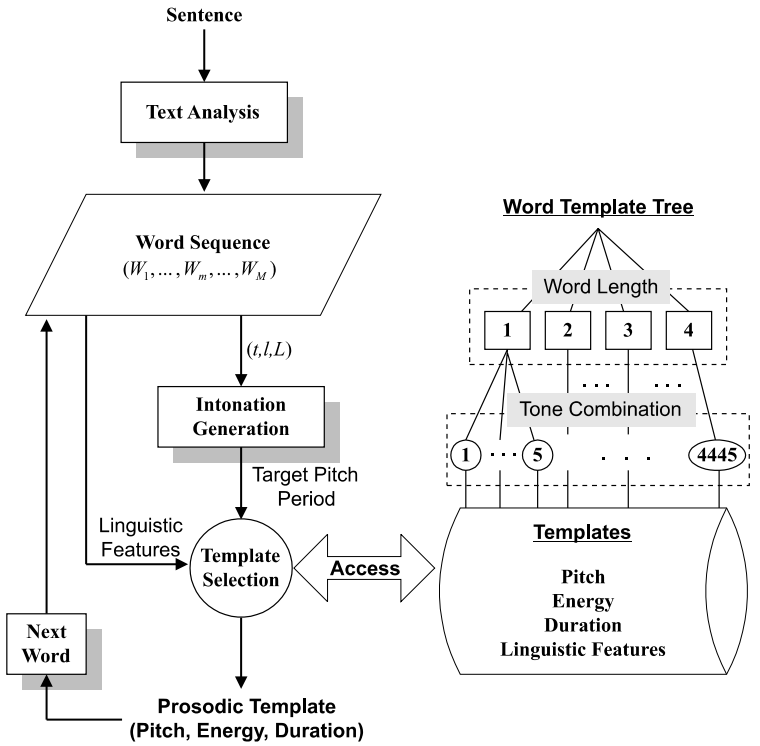

Fig. 6. Generation of the word-prosody templates for a sentence.

to the first tone in a word, by providing the average pitch period of the first syllable in the word. In Mandarin speech, the tones are not only different in the shapes of pitch contours but also in the average pitch periods. So the tone is one of the parameters in the generation of sentence intonation.

For the mth word of the word sequence in Fig. 6, the triple (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 mth word is obtained by the following equation:

$$\mathrm{TP}_m(t,l,L) = p_t \cdot r(l,L), \tag{9}$$

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

$$r(l,L) = r_0 + \left(\frac{2}{1+v^{((L+1)/2)-l}}\right)\cdot(1-r_0). \tag{10}$$

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. A plot of the ratio function with respect to syllable position is shown in Fig. 7, which indicates that a larger value of v leads to a larger change of the ratio function. 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.

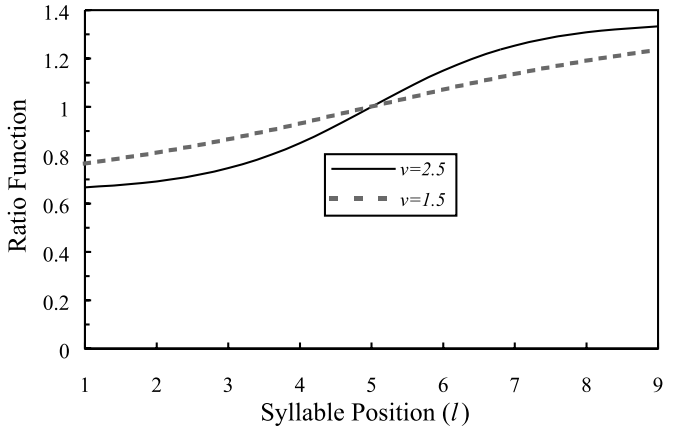

Fig. 7. A plot of the ratio function with respect to syllable position for $v = 1.5$ and 2.5 under $L = 9$ and $r_0 = 0.65$.

4.2.2. Template selection module

For the mth 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:

$$\begin{aligned}\mathrm{WT} = \{t_i \mid\ & |\mathrm{TP}_m - a_{0_i}| < \mathrm{TH},\\ & 1 \leqslant i \leqslant \text{number of all templates}\},\end{aligned} \tag{11}$$

where a_{0_i} is the component of the pitch vector of the first syllable in the ith 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

$$\begin{aligned}&\mathrm{WTC}(W_m, t_n)\\ &= D_0(\mathrm{POS}_m, \mathrm{POS}_n) + \sum_{i=1}^{|W_m|}\{D_0(\mathrm{IG}_{mi}, \mathrm{IG}_{ni})\\ &\quad + D_0(\mathrm{FG}_{mi}, \mathrm{FG}_{ni}) + D_0(I_{mi}, I_{ni})\\ &\quad + D_0(F_{mi}, F_{ni}) + D_0(\mathrm{Big5}_{mi}, \mathrm{Big5}_{ni})\}.\end{aligned} \tag{12}$$

The notations in this equation are described as follows: t_n is the word template in WT, $D_0(x, y)$ the distance measure between x and y defined as

$$D_0(x, y) = \begin{cases} 0 & \text{if } x = y,\\ 1 & \text{if } x \neq y,\end{cases} \tag{13}$$

POS the POS of the word, $|W_m|$ the word length of W_m, IG_{mi} and FG_{mi} the group numbers of the initial and the final parts of the ith syllable in W_m, respectively, I_{mi} and F_{mi} the codes of the initial and the final parts of the ith syllable in W_m, respectively, and $\mathrm{Big5}_{mi}$ is the Big5 code of the ith 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, $1 \leqslant m \leqslant M$.

Since the speech database was not well designed, some tone combinations for 3-, 4-syllable and longer words might not occur. In this case, the word templates of monosyllable and 2-syllable word are used to obtain the target prosodic features as follows:

- For 3-syllable words, the target prosodic features are obtained from a 2-syllable word followed by a monosyllable word.
- For 4-syllable words, the target prosodic features are obtained from two 2-syllable words.
- For 5-syllable and longer words, the target prosodic features are obtained from five monosyllable words.

5. Experiments and results

In our system, a continuous speech database established by the Telecommunication Laboratories, Chunghwa Telecommunication, Taiwan, containing 655 reading utterances was used to construct the synthesis unit inventory and 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.

5.1. Syllable occurrence

In this database, there are 35243 speech units which comprise 1313 different syllables. The histograms of syllable occurrence and the relative cumulative frequency are plotted in Figs. 8 and 9, respectively. It can be seen that half of the syllables have syllable occurrences less than 8, which are seldom used syllables. The percentages of syllables having 1-, 2- and 3-syllable occurrences are 18.9%, 10.4% and 6.7%, respectively. On the other hand,

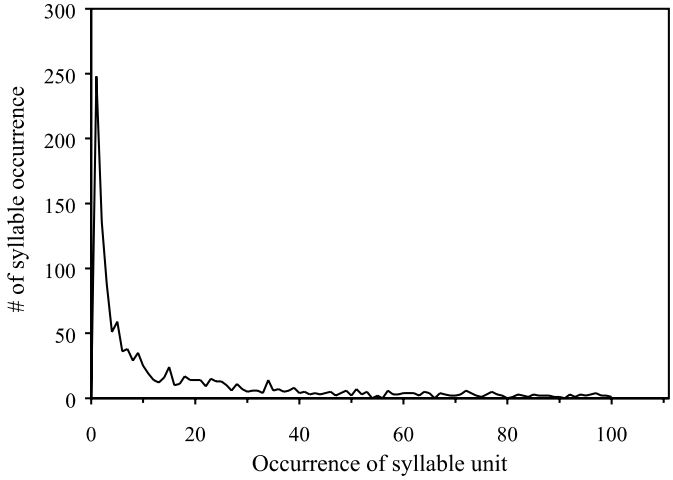

Fig. 8. Occurrence distribution of syllable unit.

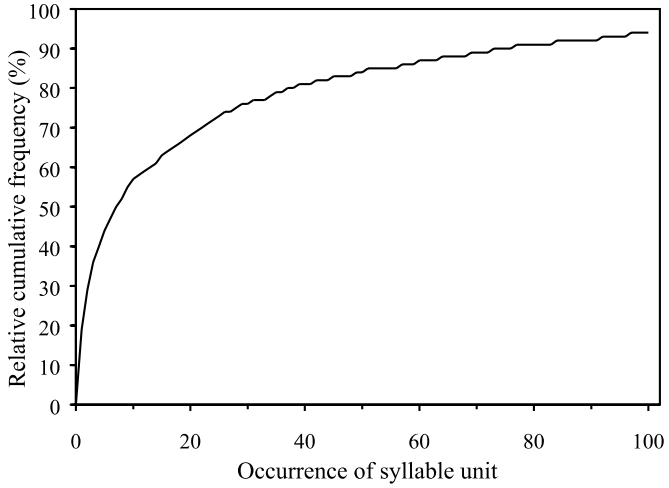

Fig. 9. Relative cumulative frequency of occurrence of syllable unit.

79 syllables have syllable occurrences more than 100. The top-2 syllable occurrences are 1135 and 751 corresponding to the syllables "de-5" and "sh-4", respectively.

5.2. Unit selection

In this experiment, a set of 1313 synthesis units was first manually and carefully selected by an experienced person which is called a manual unit set (MUS). The MUS was manually obtained with the following four criteria kept in mind: (1) only the units with good speech quality were taken into account, (2) the units with too low/high F_0 were discarded, (3) the units with short/long final part duration were discarded and (4) the units with small/large intensity were discarded. The priority for each criteria is (1) > (2) > (3) > (4). Next, the effects of the four features using in the syllable cost are investigated. The results are shown in Table 6. The term "matching rate" is defined as the hit rate of synthesis unit belonging to MUS. From this table, the matching rates for using pitch vector, intensity, duration and LSP frequencies are 45.3%, 43.1%, 41.3% and 40.6%, respectively. Pitch vector obtained the highest matching rate because it is with higher priority in manual selection. LSP frequencies are used to select the units close to their mean LSP frequencies. However, this feature does not obtain a better matching rate compared to other features. One reason is that most of the syllables have few speech units. The other reason is that the units with good speech quality might not have the LSP frequencies very close to their mean. The matching rate of combining the four features is also shown in Table 7. According to the matching rates of the four features, the weights of pitch vector, intensity, duration and LSP frequencies are $\omega^{\mathrm{P}} = 1.0$, $\omega^{\mathrm{I}} = 0.8$, $\omega^{\mathrm{D}} = 0.6$ and $\omega^{\mathrm{F}} = 0.5$, respectively. The matching rate is 48.2%, which only improves 2.9% compared to that of pitch vector. This result indicates that the four features of most of the synthesis units in the MUS are close to the corresponding mean features.

On the other hand, the effects of the syllable cost and concatenation cost are investigated. In this experiment, the syllable cost is estimated under $\omega^{\mathrm{P}} = 1.0$, $\omega^{\mathrm{I}} = 0.8$, $\omega^{\mathrm{D}} = 0.6$ and $\omega^{\mathrm{F}} = 0.5$ while the concatenation weights of the concatenation cost are $\omega_{\mathrm{l}} = 1.0$, $\omega_{\mathrm{t}} = 0.7$ and $\omega_{\mathrm{o}} = 0.5$, respectively. The matching rates of combining the syllable cost and concatenation cost are displayed in Table 8 as a function of the weight ω^{SC} for syllable cost. It can be seen that the matching rate for only using the concatenation cost ($\omega^{\mathrm{SC}} = 0$) is 42.7%, which is a little lower than that using the syllable cost ($\omega^{\mathrm{SC}} = 1$, matching rate = 48.2%). In this table, the best matching rate is 48.9% at $\omega^{\mathrm{SC}} = 0.1$. It indicates that about half of the synthesis units can be automatically obtained exactly the same with those in the MUS.

In the manual examination process, we found that most of the synthesis units of the other half are also good enough to be the synthesis units.

Table 6
Results of matching rate using different features in the syllable cost

Feature name	Pitch vector	Intensity	Duration	LSP frequencies
Matching rate	45.3%	43.1%	41.3%	40.6%

Table 7
Result of matching rate combining the features in the syllable cost

Feature name	Pitch vector	Intensity	Duration	LSP frequencies
Weight	1.0	0.8	0.6	0.5
Matching rate	48.2%			

Table 8
Matching rates as a function of the weight ω^{SC}

ω^{SC}	0	0.1	0.25	0.5	0.75	1.0
Matching rate	42.7%	48.9%	48.8%	48.4%	48.0%	48.2%

Only 63 syllables were manually replaced, that is, a replacement rate of 4.8% was obtained. The replaced syllables are generally with tight/overlapped concatenation in the speech database.

5.3. 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 9. For the male speaker, panel (a) indicates that Tones 1 and 4 have smaller values of average pitch periods (higher mean F_0) while Tones 3 and 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 little 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 dominant 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: 77 (i.e., 20000/110–20000/190), 59, 46, 77 and 38 Hz for Tones 1–5, respectively. The neutral tone has the least pitch variation (38 Hz) which is only half of that of Tones 1 and 4.

On the other hand, the intonation module proposed in the previous section employed the parameter "average pitch period p_t" for each tone. According to Table 9(a), the values of this parameter were set, respectively, as follows:

$$\begin{aligned} p_1 &= 141 \text{ for Tone 1,} \\ p_2 &= 168 \text{ for Tone 2,} \\ p_3 &= 183 \text{ for Tone 3,} \\ p_4 &= 149 \text{ for Tone 4,} \\ p_5 &= 178 \text{ for Tone 5.} \end{aligned} \tag{14}$$

5.4. Evaluation of the generated prosodic information

In this experiment, a training text was chosen for the inside test of the proposed approach. Fig. 10 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 synthesized and the original contours. It can be seen that the two contours are very similar 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 syllables match quite well with their counterparts. However, obvious deviation occurs at some syllables. A typical example is a 2-syllable word, the 17th and the 61st characters, the synthesized contours of mean energy and final part duration do

Table 9
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 (%)

(a) Tone	Average pitch period	S.D.						
1	141	23						
2	168	24						
3	185	28						
4	149	26						
5	179	27						
(b) 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

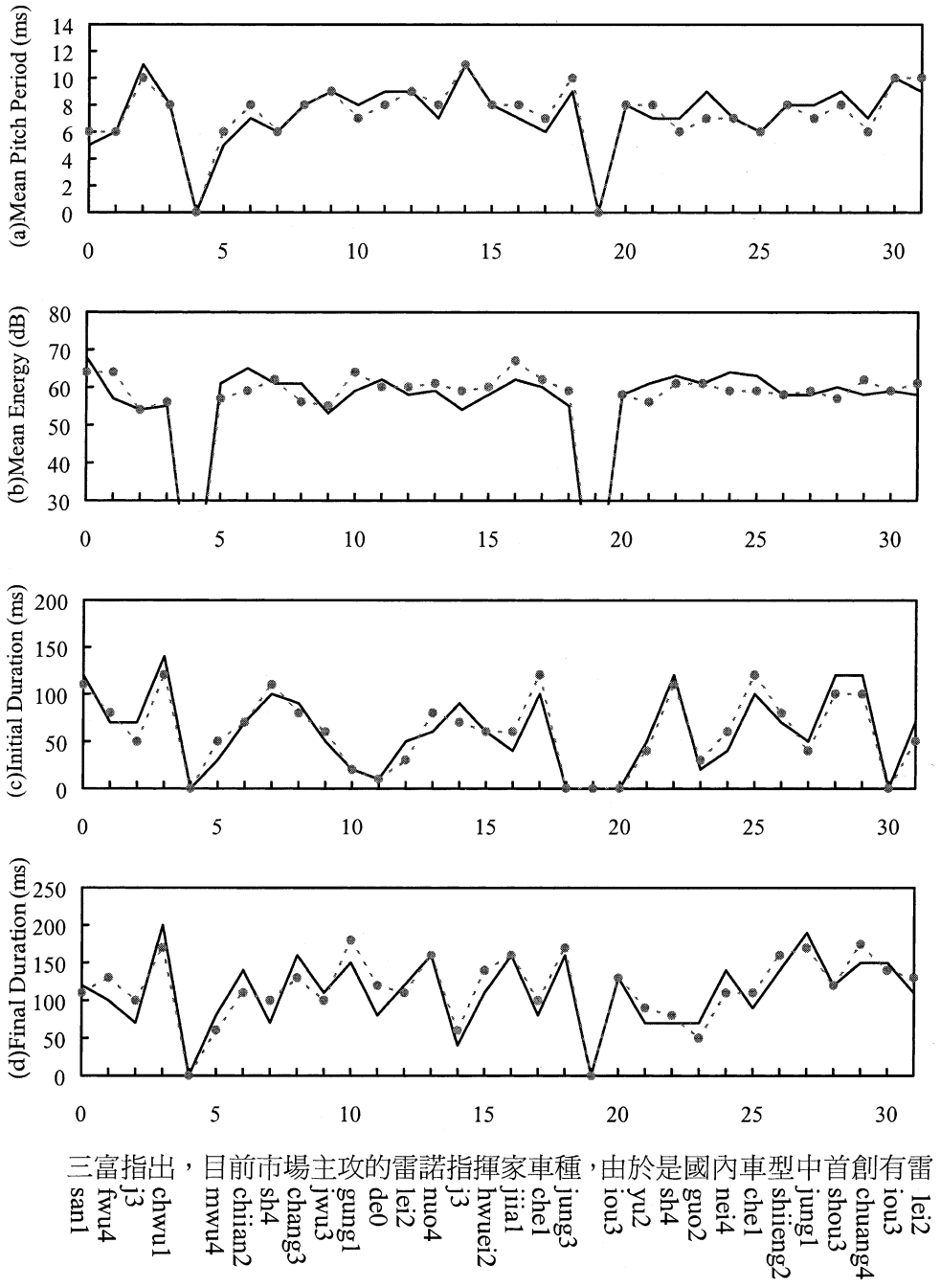

Fig. 10. Example of the original (solid lines) and the synthesized (dotted lines) prosodic parameter sequences of (a) mean pitch period, (b) average energy, (c) initial part duration, (d) final part duration. The *x*-axis represents the syllable positions corresponding to the Chinese characters in Big5.

not resemble their original ones. 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.

Some preliminary performance evaluation was conducted on this system. 20 subjects were asked to subjectively evaluate this TTS system using the following criteria:

1. *Intelligibility*. In this test, the subjects were asked to listen to a large amount of synthesized speech without prior knowledge of the content of the speech. Then, the subjects wrote down what they heard. By comparing the results with original text, the correct rate was obtained.
2. *Naturalness*. First, the subjects were asked to listen to two types of speech pronounced, respectively, by a person and by a TTS system without prosody modification. Then, the synthesized speech with prosody modification using the proposed TTS system was evaluated. For the synthesized speech, the subjects gave mean opinion scores (MOS) on a scale of 1–5, 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 of the intelligibility test is shown in Table 10. The average correct rate was 96.9%. As indicated in this table, a word with longer length was generally more intelligible since it included more semantic information. On the other hand, some words with fricative initials were inherently confusable in pronunciation, for example, the initials 'j', 'ch' and 'sh' versus the initials 'tz', 'ts' and 's', respectively. This factor largely increased the error rates. Table 11 lists the MOSs for words or sentences with different lengths. As indicated in this table, the average MOS was 3.6 for naturalness. On the contrary to the intelligibility test, the results indicate that a shorter token length obtained a higher MOS since less linguistic information was needed. Furthermore, the MOS for a short text was lower than that for the average MOS. The reason is the lack of syntactic and semantic information, which provides prosodic information in this system.

Table 10
Results for the intelligibility test

	Amount	Intelligibility
Monosyllable word	1313	93.3%
2-syllable word	200	96.0%
3-syllable word	200	98.8%
4-syllable word	200	99.2%
Sentence	100	97.2%
Average		96.9%

Table 11
Results for the naturalness test

	Amount	Naturalness
2-syllable word	200	3.8
3-syllable word	200	3.7
4-syllable word	200	3.7
Sentence	100	3.3
Short text	100	3.4
Average		3.6

6. Conclusions

In this paper, the approaches to the selection of synthesis units and prosodic information generation have been proposed for a Mandarin Chinese TTS system using a large speech database. As for synthesis unit selection, a method for pitch contour smoothing using discrete Legendre polynomials was proposed. Five procedures were proposed to select a set of high-quality synthesis units. They are: pitch-period detection and smoothing, speech unit filtering, spectral feature extraction, unit selection, and manual examination. Furthermore, four criteria were introduced to filter out unfitting speech units. Syllable and concatenation cost functions were then proposed for obtaining the synthesis units. The cost functions estimate the parameters including the prosodic features, the LSP frequencies, and types of syllable concatenation. Experimental results showed that a matching rate of 48.9% was achieved. It indicates that about half of the "best" synthesis units can be automatically obtained. On the other hand, the prosodic information was stored in a word-prosody template tree along with the linguistic fea-

tures. 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 prosodic features matched quite well with their original counterparts. Evaluation by subjective experiments also confirmed the satisfactory performance of these approaches.

References

Bigorgne, D., Boeffard, O., Cherbonnel, B., Emerard, F., Larreur, D., Le Saint-Milon, J.L., Metayer, I., Sorin, C., and White, S., 1993. Multilingual PSOLA text-to-speech system. In: Proc. ICASSP, pp. II.187–190.

Chan, N.C., Chan, C., 1992. Prosodic rules for connected Mandarin synthesis. J. Inform. Sci. Eng. 8, 261–281.

Charpentier, F.J. and Stella, M.G., 1986. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In: Proc. ICASSP, pp. 2015–2020.

Chen, S.H., Wang, Y.R., 1990. Vector quantization of pitch information in Mandarin speech. IEEE Trans. Commun. 38, 1317–1320.

Chen, S.H., Hwang, S.H., Wang, Y.R., 1998. An RNN-based prosodic information synthesizer for Mandarin text-to-speech. IEEE Trans. Speech Audio Process. 6 (3), 226–239.

Chou, F.C. and Tseng, C.Y., 1998. Corpus-based Mandarin speech synthesis with contextual syllabic units based on phonetic properties. In: Proc. ICASSP, pp. 893–896.

George, E.B., Smith, M.J.T., 1997. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal modal. IEEE Trans. Speech Audio 5 (5), 389–406.

Kawai, H., Higuchi, H., Simuzi, T. and Yamamoto, S., 1995. Development of a text-to-speech for Japanese based on waveform splicing. In: Proc. ICASSP, pp. I569–I572.

Klatt, D.H., 1987. Review of text-to-speech conversion for English. J. Acoust. Soc. Amer. 82 (3), 737–793.

Lee, L.S., Tseng, C.Y., Ouh-Young, M., 1989. The synthesis rules in a Chinese text-to-speech system. IEEE Trans. Acoust. Speech Signal Process. 37 (9), 1309–1319.

Lee, L.S., Tseng, C.Y., Hsieh, C.J., 1993. Improved tone concatenation rules in a formant-based Chinese text-to-speech system. IEEE Trans. Speech Audio Process. 1 (3), 287–294.

Quatieri, T.F., McAulay, R.J., 1992. Shape invariant time-scale and pitch modification of speech. IEEE Trans. Acoust. Speech Signal Process. 40, 497–510.

Rabiner, L.R., Schafer, R.W. (Ed.), 1978. Digital Processing of Speech Signals. Prentice-Hall, Englewood cliffs, NJ, p. 399.

Scordilis, M.S. and Gowdy, J.N., 1989. Neural network based generation of fundamental frequency contours. In: Proc. ICASSP, pp. 219–222.

Shih, C.L., Sproat, R., 1996. Issues in text-to-speech conversion for Mandarin. Comput. Linguistics Chinese Language Process. 1, 37–86.

Wang, J.F., Wu, C.H., Chang, S.H., Lee, J.Y., 1991. A hierarchical neural network model based on a C/V segmentation algorithm for isolated Mandarin speech recognition. IEEE Trans. Signal Process. 39 (9), 2141–2145.

Wu, C.H., Chen, J.H., 1997. A novel two-level method for the computation of the LSP frequencies using a decimation-in-degree algorithm. IEEE Trans. Speech Audio Process. 5 (2), 106–115.