# Automatic Selection of Synthesis Units from A Large Speech Database

*Jau-Hung Chen and Chung-Hsien Wu**

E000/CCL, Building 51, Industrial Technology Research Institute,
Hsin-Chu, Taiwan, R.O.C.
E-mail: chenjh@atc.ccl.itri.org.tw

*Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.
E-mail: chwu@csie.ncku.edu.tw

## ABSTRACT

In this paper, a novel method for the selection of synthesis unit is proposed. The monosyllables are adopted as the basic synthesis units. A set of high-quality synthesis units is selected from a large continuous speech database based on four procedures: pitch period detection and smoothing, speech unit filtering, unit selection, and manual examination. Two cost functions are proposed for obtaining the synthesis units, which minimize the inter- and intra-syllable distortion. The cost functions estimate the parameters including the prosodic features, the LSP frequencies, and types of syllable concatenation. Experimental results showed that a match rate of 48.9% was achieved. It indicates that about half of the "best" synthesis units can be automatically obtained. Also, a replacement rate of 4.8% was obtained.

## 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 [3]-[4]. A detailed overview of TTS systems for English and Chinese were introduced by Klatt [1] and Shih and Sproat [5], 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. The prosodic information generation employs the linguistic features to generate prosodic features including pitch contour, energy contour, and duration. 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 [5]. 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.

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 [1]. 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 appropriately incorporated to reduce the effect of spectral distortion [2]. Recently, the approaches to unit selection from a large speech database or using non-uniform units [6] have been appreciated and proved to obtain natural and high quality speech.

These approaches defined a cost function to select an appropriate sequence of synthesis segment.

This paper proposes a Chinese text-to-speech 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 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. However, due to the storage problem, a set of 408 syllables with the high-level tone has generally been used [4]. 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, a set of 1313 tonal monosyllables is adopted as the basic set of synthesis units, which was selected from a large continuous speech database. A novel method for synthesis unit selection is proposed, in which four procedures are proposed to select a set of high-quality synthesis units. They are pitch period detection and smoothing, speech unit filtering, unit selection, and manual examination. Fig. 1 shows the block diagram of the synthesis unit selection. Each block is described in detail in the following sections.

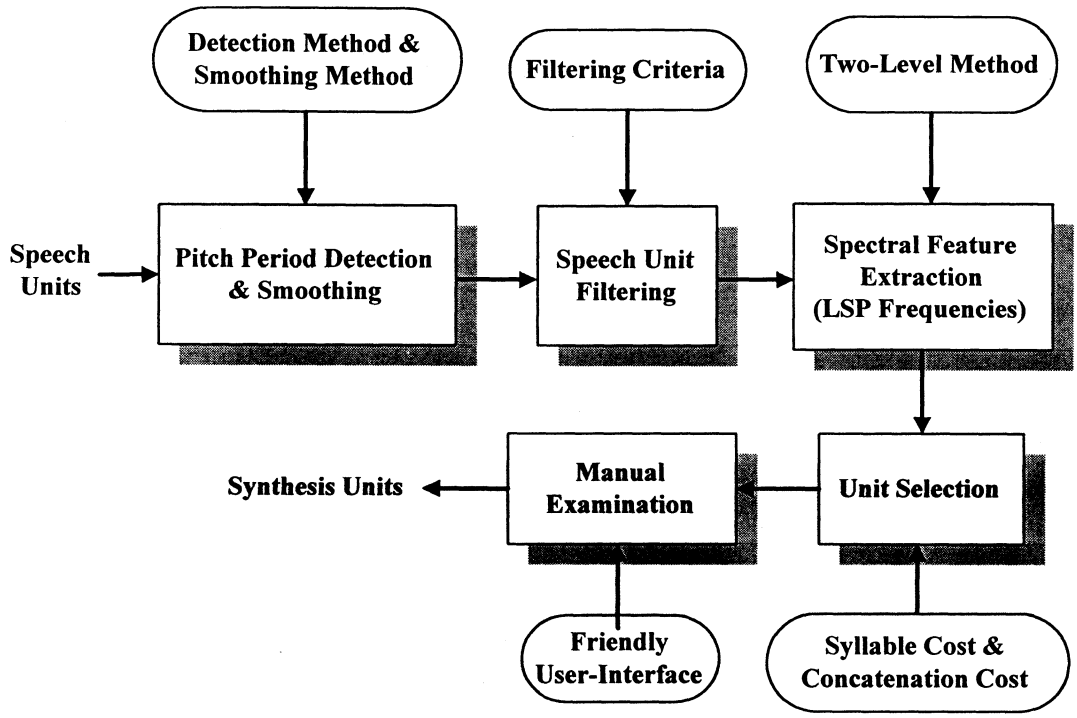


Fig. 1. Block Diagram of the Synthesis Unit Selection.

## 2. PITCH PERIOD DETECTION AND SMOOTHING

For each speech unit in the speech database, pitch-mark labeling is automatically estimated by the autocorrelation method and a modified C/V segmentation algorithm [7]. Besides, quantitative description of the pitch contours is expressed by orthonormal expansion using discrete Legendre polynomials [4]. 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 and efficient way is used to eliminate the discontinuity at both the beginning and end in a pitch contour.

Fig.2 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.

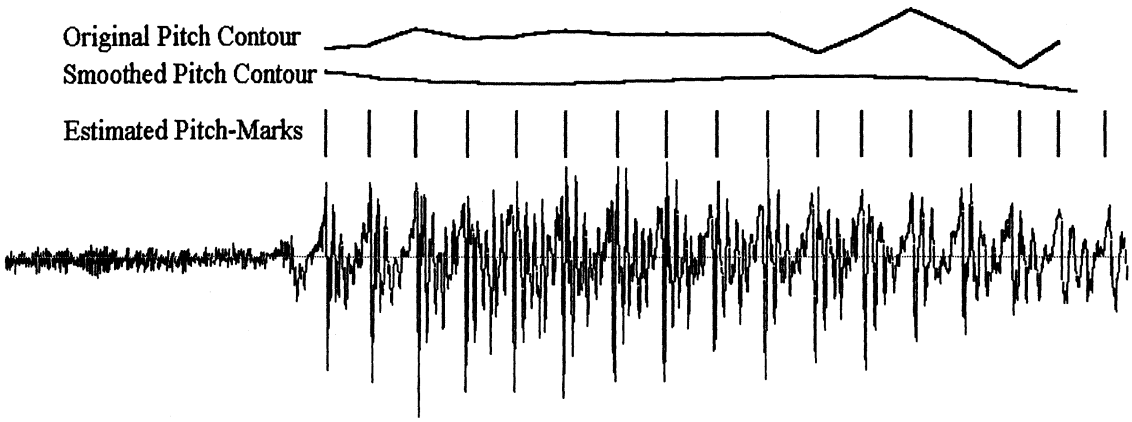


Fig. 2. An example of pitch contour smoothing using discrete Legendre polynomials.

## 3. SPEECH UNIT FILTERING

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

- Syllable duration

Generally, units with short duration imply that they are not pronounced well or completely. They 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 have few pitch-marks resulted from the unstable or short voiced parts. They are not suitable for duration or pitch modification. A threshold of 8 pitch-marks is 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.

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

Three types of inter-syllable concatenation are defined, as shown in Table 1, by investigating the inherent phonetic properties of phonemes. They are loose concatenation, tight concatenation, and

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 |

overlapped concatenation. Loose concatenation indicates few effects of inter-syllable coarticulations. So 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.

- Ranges of pitch period and syllable intensity

Syllables with extremely large/small pitch periods are poor in pitch modification. In our system, the desired average pitch periods of syllables are between $5ms$ and $11ms$ for the male speaker. Also, the desired syllable intensities are between half and twice of the average intensity for each syllable.

## 4. UNIT SELECTION

Two types of distortion measure are employed in the determination of synthesis units. One is the intra-syllable distortion (intra-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 are described as follows.

- Intra-syllable Cost

Four kinds of features are used to estimate the syllable cost: the LSP frequencies, pitch vector, mean intensity, and duration. The syllable cost is a weighted sum of the distances 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

$$SC_j(s_i) = \omega^F \cdot D_{LSP}(F_i, \overline{F}_j) + \omega^P \cdot D_P(P_i, \overline{P}_j) + \omega^I \cdot D_I(I_i, \overline{I}_j) + \omega^D \cdot D_D(D_i, \overline{D}_j) \quad (1)$$

where the notations are described as follows:

$F_i$: LSP frequencies of $s_i$;

$P_i$: Pitch vector of $s_i$;

$I_i$: Average intensity of $s_i$;

$D_i$: Duration of $s_i$;

$\overline{F}_j, \overline{P}_j, \overline{I}_j$, and $\overline{D}_j$ : Mean vectors of their corresponding features of syllable $j$.

$D_{LSP}(F_i, \overline{F}_j)$ : Distance measure for the LSP frequency, which is defined as

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

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

$D_P(P_i, \overline{P}_j)$ : Distance measure for the pitch vector, which is defined as

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

where $\sigma^{P_j}$ is the standard deviation of $P_j$.

$D_I(I_i, \overline{I}_j)$ : Distance measure for intensity, which is defined as the sum of the initial part and the final part:

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

where $\sigma^{I_j}$ is the standard deviation of $I_j$.

$D_D(D_i, \overline{D}_j)$ : Distance measure for duration, which is defined as the sum of the initial part and the final part:

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

where $\sigma^{D_j}$ is the standard deviation of $D_j$.

$\omega^F, \omega^P, \omega^I$, and $\omega^D$ : Weights of their corresponding features.

- Concatenation Cost

In our system, 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

$$CC_j(s_i) = \frac{1}{2(N-1)} \cdot \sum_{k=1, k\neq j}^{N}\left[\frac{D_l(F_i, \overline{F}_k)}{\omega_l(s_i)} + \frac{D_r(F_i, \overline{F}_k)}{\omega_r(s_i)}\right] \tag{6}$$

where $N$ is the total number of different syllables, $D_l(\bullet)$ and $D_r(\bullet)$ denote the Euclidean distance measure for spectral distortions of the left and right concatenation, respectively. It is noted that $D_l(\bullet)$ and $D_r(\bullet)$ only calculate the distances of the LSP frequencies in desired frames. $\omega_l(s_i)$ and $\omega_r(s_i)$ 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 expressed as

$$\text{Coarticulation weight} = \begin{cases} \omega_\ell & \text{for loose concatenation;} \\ \omega_t & \text{for tight concatenation;} \\ \omega_o & \text{for overlapped concatenation;} \end{cases} \quad (7)$$

where $\omega_\ell > \omega_t > \omega_o$.

Finally, a speech unit $s_i$ 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:

$$TC_j(s_i) = \omega^{SC} \cdot SC_j(s_i) + (1-\omega^{SC}) \cdot CC_j(s_i) \quad (8)$$

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

## 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.

All the above information is stored for prosody modification.

## 6. EXPERIMENTAL RESULTS

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

In this experiment, a set of 1313 synthesis units was first manually and carefully selected by an experienced person which is called the 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 pitch frequency were discarded; (3) The units with short/long final 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 intra-syllable cost are investigated. The term "match rate" is defined as the hit rate of synthesis unit belonging to MUS. From the experimental results shown in Table 2, the match 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 match 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 match 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 match rate of combining the four features is shown in Table 3. According to the match rates of the four features, the weights of pitch vector, intensity, duration,

Table 2. Results of match rate using different features in the syllable cost.

| Feature Name | Pitch Vector | Intensity | Duration | LSP Frequencies |
|---|---|---|---|---|
| Match Rate | 45.3% | 43.1% | 41.3% | 40.6% |

Table 3. Result of match 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 |
| Match Rate | 48.2% | | | |

Table 4. Match rates as a function of the weight $\omega^{SC}$.

| $\omega^{SC}$ | 0 | 0.1 | 0.25 | 0.5 | 0.75 | 1.0 |
|---|---|---|---|---|---|---|
| Match Rate | 42.7% | 48.9% | 48.8% | 48.4% | 48.0% | 48.2% |

and LSP frequencies are $\omega^P = 1.0, \omega^I = 0.8, \omega^D = 0.6$ and $\omega^F = 0.5$, respectively. The average match rate is 48.2%. This 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.

Finally, the effects of the intra-syllable cost and concatenation cost are investigated. In this experiment, the syllable cost is estimated under $\omega^P = 1.0, \omega^I = 0.8, \omega^D = 0.6$ and $\omega^F = 0.5$ while the concatenation weights of the concatenation cost are $\omega_\ell = 1.0$, $\omega_t = 0.7$, and $\omega_o = 0.5$, respectively. The match rates of combining the syllable cost and concatenation cost are displayed in Table 4 as a function of the weight $\omega^{SC}$ for syllable cost. The match rate for only using the concatenation cost is 42.7%, which is a little lower than that using the intra-syllable cost (match rate=48.2%). The best match rate is 48.9%. 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 synthesis units. Only 4.8% of the automatically chosen synthesis units were manually replaced. However, most of them were selected from a pool of poor units. These syllables are generally with tight/overlapped concatenation in the speech database.

## 7. CONCLUSIONS

The approaches to the selection of synthesis units have been proposed using a large speech database. A method for pitch contour smoothing using discrete Legendre polynomials was proposed. Also, 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 unfitting speech units out. Syllable cost and concatenation cost 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 match rate of 48.9% was achieved. It indicates that about half of the "best" synthesis units can be automatically obtained. Also, a replacement rate of 4.8% was obtained.

## REFERENCES


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

[2] D. Bigorgne, O. Boeffard, B. Cherbonnel, F. Emerard, D. Larreur, J. L. Le Saint-Milon, I. Metayer, C. Sorin, S. White "Multilingual PSOLA text-to-speech system," Proc. ICASSP, 1993, pp. II.187-II.190. Minneapolis, Minnesota, U.S.A.

[3] 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, 1(3), pp. 287-294, 1993.

[4] 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, 1998, pp. 226-239.

[5] 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.

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

[7] J. F.,Wang, C. H. Wu, S. H. Chang, and J. Y. Lee, "A hierarchical neural network model based on a C/V segmentation algorithm for isolated Mandarin speech recognition," IEEE Trans. Signal Processing, 39(9), 1991, pp. 2141-45.