Proc. Natl. Sci. Counc. ROC(A)
Vol. 21, No. 5, 1997. pp. 505-512

Bayesian Network Based Generation of Prosodic Information for Chinese Text-to-Speech Conversion

CHUNG-HSIEN WU AND JAU-HUNG CHEN

Institute of Information Engineering
National Cheng Kung University
Tainan, Taiwan, R.O.C.

(Received January 7, 1997; Accepted March 20, 1997)

ABSTRACT

In this paper, a novel approach based on Bayesian networks to the generation of prosodic information is proposed. A set of 1410 Mandarin syllables is adopted as the basic synthesis units. To enhance the traditional rule-based approach to the generation of prosodic information, the Bayesian network is employed to model the relation between the prosodic information and the linguistic features. This network is trained with a set of 112 phonetically balanced sentences and 250 sentences selected from newspapers and textbooks. Given a Chinese character sequence, the Bayesian network can provide appropriate prosodic information including pitch contour, syllable intensity, syllable duration and pause duration. Furthermore, pitch contour modification is achieved by modifying the waveform output using the pitch-synchronous overlap-and-add (PSOLA) method. The synthesized speech has been tested on 20 subjects. The results indicated that the average correct rate was 97.0% for intelligibility, and that the mean opinion score (MOS) was 3.8 for naturalness.

Key Words: text-to-speech conversion, prosodic information, pitch contour, Bayesian network

I. Introduction

Recently, many studies have focused on Text-to-Speech (TTS) systems for different languages (Klatt, 1987; Charpentier and Stella, 1986; Chen *et al.*, 1992; Hwang and Chen, 1992, 1994, 1995; Bigorgne *et al.*, 1993; Hwang, 1996). Also, TTS systems and synthesis technology for Chinese language have been developed in the last decade (Zhang, 1986; Yang and Xu, 1988; Lee *et al.*, 1989, 1993; Chan and Chan, 1992; Choi *et al.*, 1994). Aside from examination of the inherently different characteristics of each language, much effort has been focused on the rule-based approach to prosody modification, as evidenced by the vast amount of published literature. These phonological rules are invoked to imitate the pronunciation of human beings. The derivation of these phonological rules, however, is laborious, time-wasting and tedious. Furthermore, because the phonological characteristics are interactively affected by many various linguistic features, 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. For Mandarin text-to-speech systems, Hwang and Chen (1994) proposed a multi-layer perceptions (MLP) approach trained with a backpropagation (BP) algorithm to generate the pitch information. Also, they proposed a multirate recurrent neural network (MRNN) to simulate human pronunciation to explore the hidden pronunciation states embedded in an input text (Hwang and Chen, 1995). The MRNN was trained either by a simple recurrent neural net or by the BP algorithm. However, the BP algorithm of error gradients exhibits a number of serious problems in training feedforward neural networks. The user is required to select three arbitrary coefficients: the learning rate, momentum, and the number of hidden nodes. An unfortunate choice can cause slow convergence. Furthermore, the network can become trapped in a local minimum of the error function, thus arriving at an unacceptable solution when a much better one exists. Furthermore, training an MRNN with the BP algorithm requires a great deal of computation and memory, and leads to the same problems as in feedforward networks. On the other hand, most input features in these approaches were represented by a string of binary codes. For example, the 3-component input feature has been encoded as three discrete types, (1, 0, 0), (0, 1, 0) and (0, 0, 1). In such simple encoding scheme, it is difficult to represent the precise distance/similarity between the input vectors.

In this paper a Bayesian network is used to model the relationship between linguistic features and prosodic information. This network, a statistical model based on Bayes' theorem, has the following advantages compared to the above approaches. (1) The training process is faster than that using BP algorithm. (2) The Bayesian network serves as a rule-based generator. It can automatically store "rule patterns" via weights. For an unseen input vector, this network serves as a classifier to give a statistically optimal output.

On the other hand, the set of basic synthesis units to be concatenated, such as phonemes, semisyllables, diphones, syllables, etc., also plays a key-role in a TTS system. An important characteristic of Mandarin Chinese is that it is a tonal language based on monosyllables. There are five basic tones: 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 1410. 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 (Lee *et al.*, 1989, 1993). Such an approach might obtain less satisfactory results for the intelligibility test because substantial changes in the tonal manifestations of a syllable depend on the context. In this paper, the set of 1410 Mandarin Chinese monosyllables is adopted as the basic set of synthesis units. The time-domain/waveform pitch-synchronous overlap-and-add (PSOLA) method is employed for modification of prosodic information (Charpentier *et al.*, 1986; Bigorgne *et al.*, 1993).

The rest of this paper is organized as follows. A brief description of the whole system structure is presented in Section II. The Bayesian network is described in Section III. Sections IV and V introduce the processes of generation and modification of prosodic information. The performance of the system is evaluated and discussed in Section VI. Some conclusions are given in the last section.

II. System Description

A block diagram of this text-to-speech system is shown in Fig. 1 and described in detail in the following.

(1) Text analysis: A preliminary text analysis is first used to identify punctuation, numerals and Chinese characters. Meanwhile, some contextual features, such as the phonetic structure and syntactical structure, are extracted. Also, a Chinese word dictionary of about 80000 entries is available, in which the words are organized in a tree-like structure to improve the search speed. In the word segmentation process, a Chinese word segmentation module is invoked to segment a character sequence into a word sequence using the word dictionary. In the word formation process, a monosyllable word not belonging to any word is combined with its preceding word to give the final word sequence.

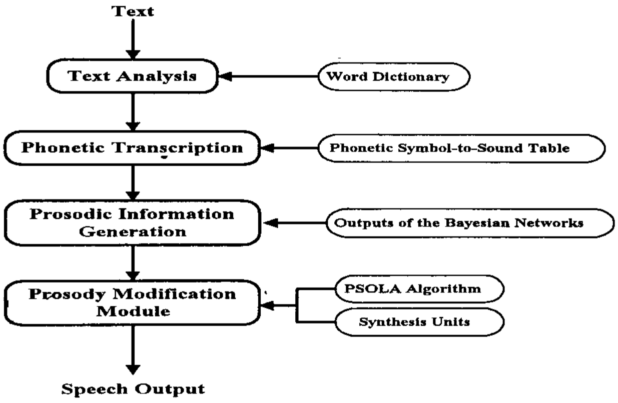

Text
Text Analysis
Word Dictionary
Phonetic Transcription
Phonetic Symbol-to-Sound Table
Prosodic Information Generation
Outputs of the Bayesian Networks
Prosody Modification Module
PSOLA Algorithm
Synthesis Units
Speech Output

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

(2) Phonetic transcription: The phonetic transcription for each character is obtained by referring to a phonetic symbol-to-sound table.

(3) Prosodic information generator: A Bayesian network is employed to model the relation between the linguistic features extracted from the input text and the fluctuation of the prosodic information. This network is trained with a set of sentence utterances and provides appropriate prosodic information to adjust the synthesis speech units.

(4) Prosody modification module: By analyzing the prosodic behavior under various concatenation conditions from sentence utterances in a database, some representative prosodic patterns are generated and chosen to represent the prosodic information in Mandarin Chinese. Using the outputs form the prosodic information generator, prosody modification based on the PSOLA approach is carried out to produce synthesized speech.

III. Bayesian Network

The Bayesian network, a statistical model based on the Bayes' theorem, is proposed to generate prosodic information. In this system, the prosodic information includes the pitch contour, syllable intensity, syllable duration and pause duration. In the training process, the parameters of the model are extracted from a set of training sentence utterances. The adjustment rules of the prosodic information can be automatically learned and stored in the Bayesian network. The structure of the network includes four layers: the input layer, the Gaussian layer, the mixture layer and the *a posteriori* layer, as shown in Fig. 2.

The input layer is composed of linguistic features extracted from the text. An adaptive weight is associated with each input to the Gaussian layer connections. The Gaussian layer functions as a vector quantizer, and each Gaussian node is assigned to a particular codeword determined by the K-means algorithm. Each mixture node represents one of the reference patterns; it accumulates the weighted outputs of the Gaussian nodes in order to respond to the similarity between the input vector and one reference pattern. For an input vector $\mathbf{X}_q$, each Gaussian node of the k-th Gaussian network calculates the conditional probability with respect to codeword G_j^k as

$$P(X_q|G_j^k) = N[\mathbf{X}_q, \boldsymbol{\mu}_j^k, \mathbf{V}_j^k], \tag{1}$$

in which N represents a multivariate Gaussian density function, $\boldsymbol{\mu}_j^k$ is the mean vector of codeword G_j^k, and $\mathbf{V}_j^k$ is the covariance matrix of codeword G_j^k. In practice, the components of input vector $\mathbf{X}_q$ are essentially uncorrelated. Thus, $\mathbf{V}_j^k$ becomes a diagonal covariance matrix; Eq. (1) is expressed simply as

$$N[\mathbf{X}_q, \boldsymbol{\mu}_j^k, \mathbf{V}_j^k] = \frac{1}{(2\pi)^{\mathrm{D}/2}\left(\prod_{d=1}^{\mathrm{D}} \sigma_{jd}^{k2}\right)^{1/2}} \exp\left(-\sum_{d=1}^{\mathrm{D}} \frac{(x_{qd}-\mu_{jd}^k)^2}{2\sigma_{jd}^{k2}}\right), \tag{2}$$

in which $\mathbf{D}$ is the dimension of the input vector, μ_{jd}^k is the d-th component of $\boldsymbol{\mu}_j^k$, and σ_{jd}^k is the d-th covariance of $\mathbf{V}_j^k$. The output of the i-th mixture node is the summation of the weighted outputs of Gaussian node, expressed as

$$P(X_q|C_i^k) = \sum_{j=1}^{M} \omega_{ij} \times P(X_q|G_j^k), \tag{3}$$

in which C_i^k is the i-th mixture node, G_j^k is the j-th Gaussian node, M is the codebook size of each Bayesian network and ω_{ij} is the weight of the j-th Gaussian node to the i-th mixture node in the Bayesian network. Explicitly,

$$\omega_{ij} = \frac{\left|\{X_q | X_q \in (C_i^k \cap G_j^k)\}\right|}{\left|\{X_q | X_q \in C_i^k\}\right|}. \tag{4}$$

In Eq. (4), the numerator is the number of vectors belonging to C_i^k and G_j^k. The denominator is the number of vectors corresponding to C_i^k. Then, the output probability of every node in the *a posteriori* layer can be calculated as follows:

$$p(C_i^k|X_q) = \frac{p(X_q|C_i^k)p(C_i^k)}{p(X_q)}. \tag{5}$$

The node with the maximum probability in the *a posteriori* layer indicates the appropriate prosodic information.

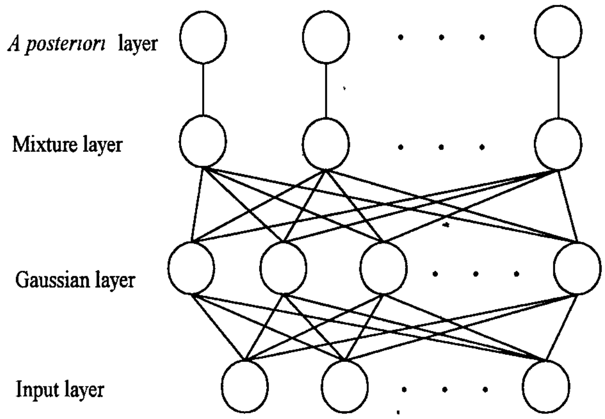


Fig. 2. The structure of the Bayesian network.

IV. Prosodic Information Generator

There are four prosodic information generators in our system. They are described as follows.

1. Pitch Contour Information Generator

Mandarin Chinese is a tonal language with four lexical tones and one neutral tone, and each tone can be described in terms of its pitch contour. It has been known for some time in linguistics that the variation of the tone/pitch contour of a Mandarin syllable is usually affected by the tones or the coarticulatory effects between adjacent syllables. This variation often causes substantial changes in the pitch contour. Therefore, it is necessary to subdivide each tone into more precise ingredients. After performing vector quantization on the set of syllables segmented from the training sentences, 12 representative pitch contour patterns are derived to represent the variation of the five tones. They are plotted in Fig. 3 and denoted a 1-a and 1-b for Tone 1; 2-a and 2-b for Tone 2; 3-a, 3-b and 3-c for Tone 3; 4-a, 4-b and 4-c for Tone 4; 5-a and 5-b for Tone 5.

On the other hand, quantitative description of the

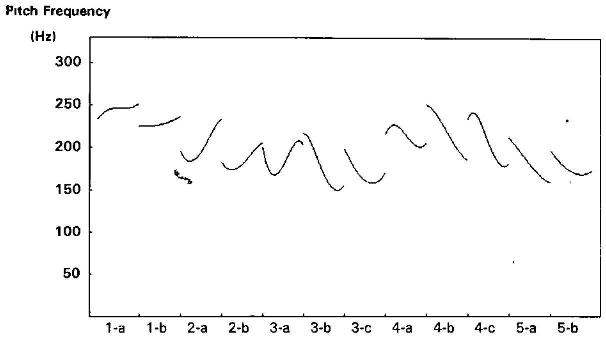


Fig. 3. The representative pitch contours.

pitch contours is expressed using orthonormal expansion with discrete Legendre polynomials (Chen and Wang, 1990; Zhang, 1986; Yang and Xu, 1988; Lee *et al.*, 1989, 1993; Chan and Chan, 1992). Given a syllable with a length of N+1 frames, suppose the pitch frequency value of the n-th frame is represented by $P(\frac{n}{N})$ and normalized in the interval [0, 1]. Then, $P(\frac{n}{N})$ can be approximately expressed in terms of the first four discrete Legnedre polynomials:

$$P(\tfrac{n}{N}) = \sum_{i=0}^{3} a_i \Phi_i(\tfrac{n}{N}), \qquad 0 \le n \le N, \tag{6}$$

where

$$\Phi_0(\tfrac{n}{N}) = 1 \tag{7}$$

$$\Phi_1(\tfrac{n}{N}) = (\frac{12N}{N+2})^{1/2}[(\tfrac{n}{N}) - \tfrac{1}{2}] \tag{8}$$

$$\Phi_2(\tfrac{n}{N}) = [\frac{180N^3}{(N-1)(N+2)(N+3)}]^{1/2}[(\tfrac{n}{N})^2 - (\tfrac{n}{N}) + \frac{N-1}{6N}] \tag{9}$$

$$\Phi_3(\tfrac{n}{N}) = [\frac{2800N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}]^{1/2} \times [(\tfrac{n}{N})^3 - \tfrac{3}{2}(\tfrac{n}{N})^2 + \frac{6N^2-3N+2}{10N^2}(\tfrac{n}{N}) - \frac{(N-1)(N-2)}{20N^2}] \tag{10}$$

and

$$a_i = \frac{1}{N+1}\sum_{n=0}^{N} p(\tfrac{n}{N})\Phi_i(\tfrac{n}{N}), \qquad 0 \le i \le 3. \tag{11}$$

In other words, each pitch contour of a syllable is represented by a four-dimensional vector, $\vec{a} = (a_0, a_1, a_2, a_3)$. This quantitative description and Euclidean distance measure are employed in the vector quantization.

The training process is as follows. For each syllable with tone k in the training set, ten feature components are extracted from the input text:

(1) Four feature components represent the pitch contour of the preceding syllable and are denoted by $(a_{10}, a_{11}, a_{12}, a_{13})$.

(2) Four feature components represent the pitch contour of the succeeding syllable and are denoted by $(a_{20}, a_{21}, a_{22}, a_{23})$.

(3) Another two feature components represent the positions of the syllable in a sentence and in a word; their lengths are normalized in the interval [0, 1] and denoted by s and ω, respectively.

The combined feature vector $(a_{10}, a_{11}, a_{12}, a_{13}, a_{20}, a_{21}, a_{22}, a_{23}, s, \omega)$ is used as the input vector of the Bayesian network. The nodes G_j^k in the Gaussian layer can be constructed by the K-means algorithm. The nodes C_i^k in the mixture layer are the representative pitch contour patterns. The weights ω_{ij} between the Gaussian layer and mixture layer can be obtained by Eq. (4). On the other hand, one of the C_i^k's is considered to be the desired (target) class of pitch contour patterns for the input syllable. The desired class can be obtained in advance by choosing the one with the smallest Euclidean distance:

$$C_{i^*}^k = \arg\min_{1 \le i \le N} \left\| \vec{C}_i^k - \vec{b} \right\|, \tag{12}$$

where $\vec{b}$ is the feature vector of the training syllable.

In the synthesis process, the input vector is extracted from the input text as in the training process. After calculating Eqs. (2), (3) and (5), the pitch contour pattern, which is the most probable one for the input syllable, is obtained and then fed to the pitch modification module.

It is believed that Part-of-Speech (POS) features are very helpful in generating good prosody of sentences (Hwang, 1996). However, they are excluded in this paper because they can not provide appropriate prosodic information without any semantic information from the context. Given a sentence with a structure of "Noun Verb Noun," one can utter the sentence in different ways by emphasizing any of the POSs. Therefore, a POS should be further labeled by a semantic parser. However, a good semantic parser is not available at the present time.

In general, the procedures for training and generating the syllable intensity, syllable duration and pause duration information using the Bayesian network are similar to those for pitch contours. To be brief, we will only list the differences in the following.

2. Syllable Intensity Information Generator

(1) The syllable intensity is quantized into 8 levels.
(2) The input vector includes 10 feature components; four feature components indicate tones and phonemes of the two preceding adjacent syllables, four feature components indicate tones and phonemes of the two succeeding syllables, and two other feature components indicate the positions of the syllable in the sentence and in the word. Also, their lengths are normalized so that they will lie between 0 and 1.
(3) A syllable intensity Bayesian network with 8 outputs is established and used to choose the intensity level for each syllable.

3. Syllable Duration Information Generator

(1) The syllable duration is quantized into 5 levels.
(2) The input vector is the same as that of the intensity information generator.
(3) A syllable duration Bayesian network with 5 outputs is established and used to choose the syllable duration for each syllable.

4. Pause Duration Information Generator

(1) The pause duration for each punctuation is quantized into 2 levels.
(2) A pause duration Bayesian network with 2 outputs is established for each of the 10 punctuation.
(3) The input vector is the same as that of the syllable intensity information generator.

V. Prosody Modification Module

Prosody adjustment is based on the pitch-synchronous overlap-and-add (PSOLA) approach in the time domain, which is capable of providing good quality prosody modification of natural speech (Charpentier and Stella, 1986; Bigorgne *et al.*, 1993). The synthesized signal $S'(n)$ is obtained from the digitized signal waveform $S(n)$ by

$$S'(n) = \frac{\alpha \sum_q S_{m(q)}(t_{m(q)} + n - t'_q) h_q(t'_q - n)}{\sum_q h_q^2(t'_q - n)}, \qquad (13)$$

where $S_m(n)$ is a sequence of short-term overlapping signals, α is a scale of the amplitude, $h_q(n)$ represents a Hanning windows centered around the time origin n=0, t_m is a sequence of pitch-marks on the voiced portions of $S_m(n)$, and t'_q is a new sequence of pitch-marks on the voiced portions of $S'_q(n)$.

The correspondence between the analysis short-term signals and the synthesis short-term signals is specified by the time-warping function $t_{m(q)}$. The principle of this synthesis scheme is equivalent to minimizing the quadratic error between the spectra of the analysis short-term signals and the corresponding short-term spectra of the synthetic signals (Charpentier and Stella, 1986). Modification of the pitch contour and syllable intensity is achieved by giving a new sequence of pitch-marks and adjusting the value of α, respectively. In the process of modifying the syllable duration, stationary frames are first calculated for insertion or deletion. The frame with the least accumulative spectral distortion is referred to as the stationary frame. It can be obtained by calculating the Euclidean distances of each frame to its adjacent frames:

$$S_i = \sum_{n=-2}^{2} d_{i,i+n}, \qquad \mathbf{(14)}$$

where

$$d_{ij} = \sum_{n=1}^{p} [c_j(n) - c_i(n)]^2 \qquad \mathbf{(15)}$$

is a p-th order cepstral distance between frame i and frame j, and $c_i(n)$ is the n-th cepstral coefficient of frame i. Finally, adjustment of the pause duration is achieved by simply inserting a duration of silence into each speech segment.

VI. Performance Evaluation

Two sets of utterances were recorded by a native female speaker with intonation, naturalness and fluency kept in mind. The speech signals were first digitized by a 16-bit A/D converter at a 11.025 kHz sampling rate. The first set consisted of 1410 syllables used as the basic synthesis units. The second set included 112 phonetically balanced sentences and 250 sentences selected from newspapers and textbooks. The distribution of possible tone concatenations for each tone is listed in Table 1. The syllable segmentation and pitch-mark labeling were first automatically evaluated by a modified C/V segmentation algorithm (Wang *et al.*, 1991). They were then manually examined and adjusted using a friendly supervised interface.

This text-to-speech system was implemented on a PC/AT 486 computer. 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

Table 1. Distribution of Tone Concatenations for Adjacent Syllables for Each Tone (Tone 1 to Tone 5) in the Training Database

(a) Tone 1

		Tone of the following syllable				
		1	2	3	4	5
Tone of the preceding syllable	1	5.1%	7.5%	4.8%	6.1%	3.7%
	2	5.8%	4.4%	4.1%	5.1%	1.0%
	3	4.1%	3.7%	3.7%	6.1%	1.0%
	4	5.4%	8.2%	5.1%	6.1%	2.0%
	5	0.7%	1.0%	2.7%	2.0%	0.3%

(b) Tone 2

		Tone of the following syllable				
		1	2	3	4	5
Tone of the preceding syllable	1	5.1%	6.5%	2.8%	7.7%	2.3%
	2	3.1%	3.4%	6.0%	6.5%	1.7%
	3	5.1%	3.7%	2.6%	6.0%	1.4%
	4	4.3%	6.8%	7.7%	6.5%	1.7%
	5	1.4%	2.3%	2.3%	2.8%	0.3%

(c) Tone 3

		Tone of the following syllable				
		1	2	3	4	5
Tone of the preceding syllable	1	7.7%	5.6%	0.9%	9.4%	3.4%
	2	6.4%	9.0%	1.3%	15.4%	4.7%
	3	0.0%	0.9%	0.0%	1.7%	0.0%
	4	6.9%	6.6%	5.0%	9.2%	2.4%
	5	2.1%	1.7%	0.0%	1.7%	0.9%

(d) Tone 4

		Tone of the following syllable				
		1	2	3	4	5
Tone of the preceding syllable	1	4.5%	3.1%	2.8%	7.3%	1.2%
	2	4.7%	5.9%	4.0%	6.2%	2.1%
	3	2.6%	6.6%	2.6%	7.6%	2.1%
	4	6.9%	6.6%	5.0%	9.2%	2.4%
	5	2.1%	1.7%	0.7%	1.4%	0.5%

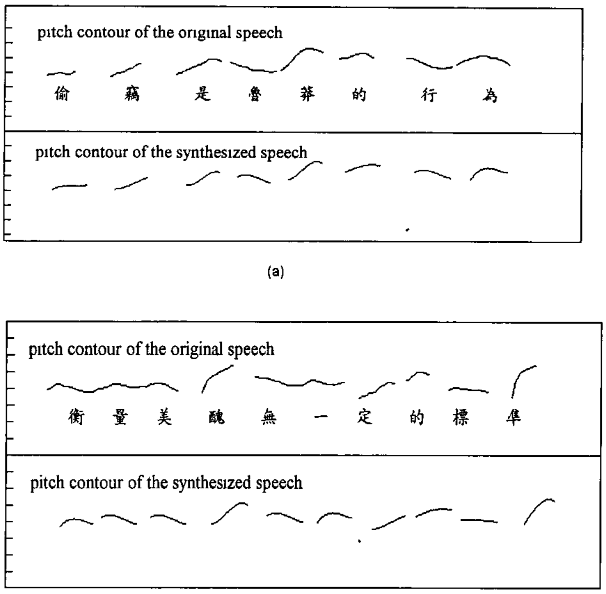
(e) Tone 5

		Tone of the following syllable				
		1	2	3	4	5
Tone of the preceding syllable	1	2.0%	11.2%	3.1%	5.1%	0.0%
	2	4.1%	7.1%	5.1%	5.1%	2.0%
	3	6.1%	6.1%	2.0%	9.2%	0.0%
	4	10.2%	7.1%	5.1%	7.1%	0.0%
	5	0.0%	1.0%	0.0%	1.0%	0.0%

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 prosodic modification. Then, the synthesized speech with prosodic modification using the proposed TTS system was 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.

Two examples of pitch contours of the original speech and synthesized speech are shown in Fig. 4. The evaluation for the intelligibility test is shown in Table 2. The average correct rate was 97.0%. 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' vs. the initials 'tz', 'ts', and 's', respectively. This factor largely increased the error rates. Table 3 lists the MOS's for words or sentences with different lengths. As indicated in this table, the average MOS was 3.8 for naturalness. Contrary to the intelligibility test, the results indicate that a shorter token length obtained a higher MOS since less linguistic information was needed. Furthermore,



(a)

(b)

Fig. 4. Two examples of pitch contours of original and synthesized speech. (a) Sentence "tou-1 chie-4 sh-4 lwu-3 mang-3 de-5 shieng-2 wei-2." (b) Sentence "heng-2 liang-2 mei-3 chou-3 wu-2 yi-2 dieng-4 de-5 biau-1 jweng-3." (The translation symbols are in Mandarin Phonetic Symbols II, and the numbers denote the tones of their corresponding syllables.)

Table 2. Results for the Intelligibility Test

	Amount	Intelligibility
Monosyllable word	1410	92.8%
2-syllable word	200	95.8%
3-syllable word	200	98.9%
4-syllable word	200	99.4%
Sentence	100	98.2%
Average		97.0%

Table 3. Results for the Naturalness Test

	Amount	Intelligibility
2-syllable word	200	4.0
3-syllable word	200	3.9
4-syllable word	200	3.9
Sentence	100	3.7
Short text	100	3.6
Average		3.8

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.

VII. Conclusion

In this paper, a Bayesian network based approach to the generation of prosodic information for Chinese text-to-speech conversion has been proposed to enhance the conventional rule-based approach. This network can obtain a faster training speed and reach a statistically optimal solution. This network can automatically store "rule patterns" via weights. For the encoding of input patterns, a metric function is used in which the metric function on the vector space is treated as a real-value function on the Cartesian product. This encoding scheme is capable of giving precise and real-valued distances between input patterns. For good speech quality, a large set of 1410 Mandarin syllables is adopted as the basic synthesis units without compression. Prosodic information, including the pitch contour, syllable intensity, syllable duration and pause duration, have been generated quickly by well-trained Bayesian networks. The PSOLA method has been adopted for pitch contour modification. Evaluation by means of listening tests has confirmed the good performance of this scheme.

Acknowledgment

The authors would like to thank the National Science Council, the Republic of China, for financial support of this work under contract no. NSC 84-2622-E006-006.

References

Bigorgne, D., O. Boeffard, B. Cherbonnel, F. Emerard, D. Larreur, J. L. Le Saint-Milon, I. Metayer, C. Sorin, and S. White (1993) Multilingual PSOLA text-to-speech system. *Proc. ICASSP*, pp. II. 187-**II**.190. Minneapolis, MI, U.S.A.

Chan, N. C. and C. Chan (1992) Prosodic rules for connected mandarin synthesis. *Journal of Information Science and Engineering*, **8**, 261-281.

Charpentier, F. J. and M. G. Stella (1986) Diphone synthesis using an overlap-add technique for speech waveforms concatenation. *Proc. ICASSP*, pp. 2015-2020. Tokyo, Japan.

Chen, S. H., S. H. Hwang, and C. Y. Tsai (1992) A first study on neural net based generation of prosodic and spectral information for mandarin text-to-speech. *ICASSP*, pp. 45-48. San Francisco, CA, U.S.A.

Chen, S. H. and Y. R. Wang (1990) Vector quantization of pitch information in Mandarin speech. *IEEE Trans. On Commun.*, **38**, 1317-1320.

Choi, J., H. W. Hon, J. L. Lebrun, S. P. Lee, G. Loudon, V. H. Phan, and S. Yogananthan (1994) Yanhui, a software based high performance Mandarin text-to-speech system. *Proc. of ROCLING VII*, pp. 35-50. National Tsing Hua University, Hsinchu, Taiwan, R.O.C.

Hwang, S. H. and S. H. Chen (1992) Neural network synthesizer of pause duration for Mandarin text-to-speech. *Electronics Letters*, **28**, 720-721.

Hwang, S. H. and S. H. Chen (1994) A neural network based F0 synthesizer for Mandarin text-to-speech system. *IEE Proc. -Vis. Image Signal Process.*, **141**(6), 384-390.

Hwang, S. H. and S. H. Chen (1995) A prosodic model of Mandarin speech and its application to pitch level generation for text-to-speech. *Proc. ICASSP*, pp. 616-619. Detroit, MI. U.S.A.

Hwang, S. H. (1996) *A Study on Prosodic Information Generator for Mandarin Text-to-Speech*. Ph.D. Dissertation. National Chiao Tung University, Hsinchu, Taiwan, R.O.C.

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

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

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

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

Yang, S. and Y. Xu (1988) An acoustic-phonetic oriented system for synthesizing Chinese. *Speech Communication*, **7**, 317-325.

Zhang, J. (1986) Acoustic parameters and phonological rules of a text-to-speech system for Chinese. *Proc. ICASSP*, pp. 2023-2026. Tokyo, Japan.

中文文句翻語音中以拜氏網路爲基礎音韻訊息之產生

吳宗憲　陳昭宏

國立成功大學資訊工程研究所

摘　要

本論文中，我們提出了一利用拜氏網路以產生音韻訊息之方法。本系統採用1410個國語單音作爲語音合成單元。爲加強傳統方法中利用規則作爲音韻訊息調整之方式，本系統利用112個音韻平衡句及250個挑選之文句作爲訓練資料庫。音韻訊息包含音高走勢、音節強度、音節長度及音節間距。此外我們利用基週同步疊加演算法來調整音高走勢。在實驗方面，我們測試20個聽眾，結果顯示平均可辨度爲97.0%，而在自然度方面，平均鑑定分數（MOS）爲3.8分。