An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech

Sin-Horng Chen, *Senior Member, IEEE,* Shaw-Hwa Hwang, and Yih-Ru Wang

***Abstract*—A new RNN-based prosodic information synthesizer for Mandarin Chinese text-to-speech (TTS) is proposed in this paper. Its four-layer recurrent neural network (RNN) generates prosodic information such as syllable pitch contours, syllable energy levels, syllable initial and final durations, as well as inter-syllable pause durations. The input layer and first hidden layer operate with a word-synchronized clock to represent current-word phonologic states within the prosodic structure of text to be synthesized. The second hidden layer and output layer operate on a syllable-synchronized clock and use outputs from the preceding layers, along with additional syllable-level inputs fed directly to the second hidden layer, to generate desired prosodic parameters. The RNN was trained on a large set of actual utterances accompanied by associated texts, and can automatically learn many human-prosody phonologic rules, including the well-known Sandhi Tone 3 F0-change rule. Experimental results show that all synthesized prosodic parameter sequences matched quite well with their original counterparts, and a pitch-synchronous-overlap-add-based (PSOLA-based) Mandarin TTS system was also used for testing of our approach. While subjective tests are difficult to perform and remain to be done in the future, we have carried out informal listening tests by a significant number of native Chinese speakers and the results confirmed that all synthesized speech sounded quite natural.**

***Index Terms*—Mandarin, pitch contour, prosodic information synthesizer, recurrent neural network, text-to-speech.**

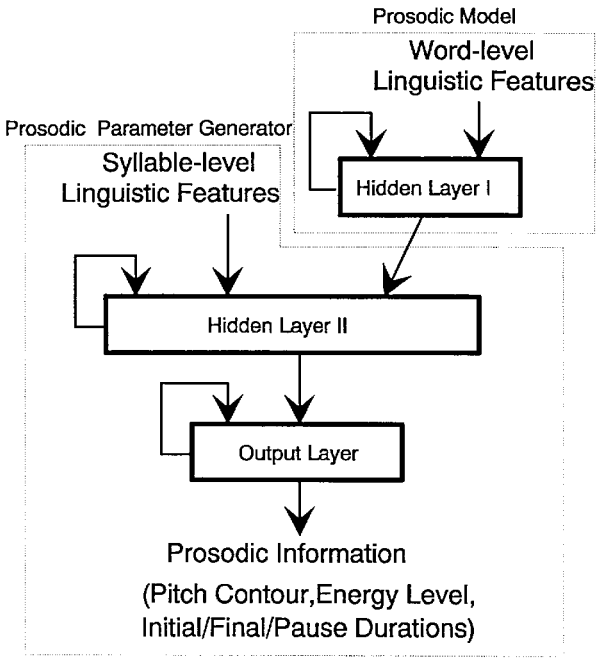

Fig. 1. Block diagram of the proposed RNN prosodic information synthesizer.

I. Introduction

In THIS paper, a new data-driven method of prosodic information synthesis for Mandarin text-to-speech (TTS) is presented. The basic idea is to use a model to explore the relationship between the prosodic phrase structure of Mandarin speech and the linguistic features of the input text for simulating human's prosody pronunciation mechanism. The model is realized by a four-layer recurrent neural network (RNN). Fig. 1 depicts the block diagram of the RNN. As Fig. 1 shows, the RNN can be functionally divided into two parts. The first part is taken as a prosodic model to explore the prosodic phrase structure of the spoken Mandarin language. It processes word-level linguistic features to track the phonologic state of the prosodic phrase structure of the utterance to be synthesized. The second part is the real prosodic information generator. It processes syllable-level linguistic features with the help of the outputs of the first part to generate all prosodic parameters needed by our Mandarin TTS system. These two parts are tightly coupled and integratedly trained using a large database to learn automatically to induce human's prosody phonologic rules. After well training, the RNN acts as a synthesizer to generate proper prosodic parameters for synthesizing natural Mandarin speech.

The main ideas of using this type of RNN to realize the prosody generation model of human are discussed as follows. We start from explaining the reasons why the first part of the RNN, which is a one-hidden-layer simple recurrent network, can be used as a prosodic model to explore the prosodic phrase structure of the input text by using only inputs of word-level linguistic features. First, because words are the smallest meaningful units of pronunciation, they should also be the basic building elements of the prosodic phrases of the spoken Mandarin language. Second, the prosodic model describing the prosodic phrase structure of a Mandarin utterance can be regarded as a model to define the relation of its constituent words; therefore, we can explore from a word sequence the prosodic phrase structure of the corresponding utterance to be

Manuscript received December 10, 1995; revised February 6, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas D. O'Shaughnessy. This work was supported by the National Science Council under Contract NSC83-0404-E-009-091.

S.-H. Chen and Y.-R. Wang are with the Department of Engineering, National Chiao Tung University, Hsinchu, Taiwan 300, R.O.C.

S.-H. Hwang is with the Department of Information Management, Ming Shin Institute of Technology, Hsinchu, Taiwan 300, R.O.C.

Publisher Item Identifier S 1063-6676(98)02897-1.

synthesized once we know the model. Third, the architecture of the first part of the RNN is similar to the simple RNN used in the studies of [1]–[5] in which the grammatical structure of a word sequence was explored via a simple task of word class prediction. So it is a dynamic system suitable for use to model the relations of words in Mandarin utterances. Based on above discussion, we believe that the first part of the proposed RNN with inputs of word-level linguistic features can function as a prosodic model. It is worth noting that, due to the following two reasons, we did not use high-level syntactical features as input features of the prosodic model in this study. First, it is generally not easy to do automatic syntactic analyses for unlimited texts of natural Chinese language. Second, the syntactic structure of a Chinese text is generally not isomorphic to the prosodic phrase structure of the corresponding Mandarin speech.

The function of the second part of the RNN prosody synthesizer is explained here in more detail. It is composed of two layers of neurons: the second hidden layer and the output layer of the RNN. Both layers have the same simple recurrent structure as the first part of the RNN to feed back all their outputs as contextual inputs to themselves. While the second part operates in the same way as the first, its functions are different owing to different driving inputs. The second hidden layer accepts two sets of inputs. One is the outputs of the first part to account for all the affections from high-level linguistic features. The other is some syllable-level linguistic features fed in directly to consider the local lexical influence. With these inputs, the second hidden layer functions as a finite state machine to model the fine (local) structure of the prosodic phrase of the input text at the current syllable. The output layer accepts the outputs of the second hidden layer to function as a predictor for the generation of all desired prosodic parameter sequences. Since all outputs are fed back as contextual inputs, the predictor is a dynamic system capable of dealing well with the temporal correlation of the output prosodic parameters, such as the declination effect on both the pitch and energy contours of declarative utterances. Besides, mutual dependencies among different types of prosodic information can be properly taken into consideration.

Lastly, the idea to derive a proper training procedure for the proposed RNN prosody synthesizer is discussed. Usually, the training procedure of a neural network-based system plays a key role to make it succeed. In the past, the training of a prosodic model for Mandarin language was practically difficult because no well-labeled training databases were available. This is mainly owing to the lack of clear and explicit definition of prosodic categories or states which construct the prosodic phrases. Thus, it is improper to train the two parts of the RNN separately. Instead, a straight forward training procedure is adopted in this study. The two parts of the RNN are tightly coupled and trained together. By directly feeding in linguistic features to the input layer and setting the prosodic parameters extracted from the training utterances as the desired output targets, these two parts can then be trained in an integrated fashion by the extended backpropagation (EBP) algorithm for recurrent neural network [6].

There are several advantages of the proposed RNN-based prosody generation method as compared to rule-based [7]–[10] and previous neural network-based [11]–[14] methods. First, the proposed method provides a total solution to the problem of prosodic information synthesis, with all prosodic parameters simultaneously generated by the compact RNN; in contrast to most previous neural network-based methods that dealt with the F0 synthesis [12]–[14] or the segmental duration synthesis [11] only. Second, only very simple inputs of word-level and syllable-level linguistic features are used. No complicated syntactic analyses are needed to extract high-level linguistic features such as major and minor phrases [12] and accent values of syllables [13]. Third, the prosodic phrase structure of the spoken Mandarin language are properly modeled and automatically trained from the real speech. There is no need to explicitly define what is a prosodic event or state in advance. It is also not necessary to manually detect either major prosodic breaks or minor prosodic breaks of the training utterances in the preprocessing stage of the training process. Fourth, all the prosody synthesis rules are embedded in the weights of the RNN and can be learned automatically without the help of any linguistic experts.

The paper is organized as follows. A general background of the prosody generation in TTS is given in Section II. It is intended to show the features shared by all languages and also those unique to Mandarin, from TTS viewpoint. The proposed method of prosodic information synthesis for Mandarin TTS is discussed in Section III. The effectiveness of the method is examined by simulations in Section IV. Some conclusions are given in the last section.

II. Background

Continuous speech contains the actual words spoken as well as suprasegmental information, such as stress, timing structure, and fundamental frequency (F0) contour patterns. This information is generally referred to as the prosody of the speech, which is affected in turn by the sentence type, the syntactical structure, the semantics, the emotional state of the speaker, etc. Without prosody, speech would be flat and toneless and would sound tedious, unpleasant, or even barely intelligible. So generating proper prosodic information is the most important issue in synthesizing natural speech in TTS systems.

Generic TTS systems need to generate F0 contours, energy contours, and word durations as well as interword pause durations. This prosodic information is usually generated according to linguistic cues extracted from the input text. Different levels of linguistic cuing, ranging from low-level lexical features, such as word phonetic structures, to high-level features, such as syntactical boundaries, can be used. Many methods for prosodic information synthesis have previously been proposed. They can be divided into two general approaches: rule-based and data-driven. In rule-based methods [7], [8], [15]–[26], input text is first analyzed to extract relevant linguistic cues. They may include lexical information such as the phonetic structures and accented word syllables, syntactical structure, intonation patterns, and declination effects of sentential utter-

ances, semantic features, etc. Phonologic rules are then used to generate the required prosodic information. Usually, the phonologic rules for synthesis are inductively inferred from observation of a large set of utterances with the help of linguists. These methods have two disadvantages. First, the rule-inference process is labor-intensive. Second, manually exploring the effect of mutual interactions among linguistic features on different levels is highly complex. As a result, it is very difficult to collect enough rules without long-term devotion to the task [7]. On the other hand, data-driven methods [11]–[14], [27]–[37] generate prosodic information from models designed to describe the relationships between linguistic features of input texts and prosodic information about the utterances to be synthesized, usually with the aid of statistical models [32]–[34] or neural networks [11]–[14], [35]–[37]. The models are trained on large sets of real utterances accompanying by associated texts. The training goals are automatic deduction of phonologic rules from the large database and implicit memorization of them in the model's parameters or the neural network's weights. During synthesis, the best combinations of prosodic information are estimated from among the models according to analysis of the linguistic features in the given input text. The primary advantage of this approach is that the phonologic rules can be automatically established from the training data set in/during the training process without the help of any linguistic expert.

Although many methods for TTS prosody generation have previously been proposed for various languages [7]–[9], [18], [25], [27], [38]–[41], it is still generally difficult to elegantly invoke high-level linguistic features in exploring the prosodic phrase structure of a spoken language for prosodic information generation. The resulting synthesized prosodic parameters are therefore inadequate for generating natural, fluent and unrestricted synthetic speeches. This is especially true of F0 synthesis because it is the most important prosodic element in determining the naturalness of synthetic speech. Recently, researchers have become aware that the fundamental problem in TTS system prosodic information synthesis is the lack of an appropriate prosodic model that describes the prosodic phrase structure of spoken language [42]–[45]. Although previous studies [44] have shown that the generally accepted prosodic phrase structure of a language is known to consist of two levels, including the intonational phrase and the intermediate phrase, its relationship to the linguistic features of the associated text is still not clearly known and needs to be explored further. Information about the prosodic phrase structure of an utterance is explicitly carried on the contours of all prosodic parameters. But, it must also be implicitly embedded in the text because it can be generated from the input text (by humans). So, a prosodic model can be generally defined as a mechanism for describing the relationship between the acoustic features extracted from the prosodic parameter contours of speech and the linguistic features extracted from the associated text. Two basic types of prosodic model can be found. One is designed to detect the prosodic phrase structure of an utterance by using some features extracted from the prosodic parameter contours [19], [45]–[53]. Its purpose is to provide either an additional score to help speech recognition [45], [50] or target

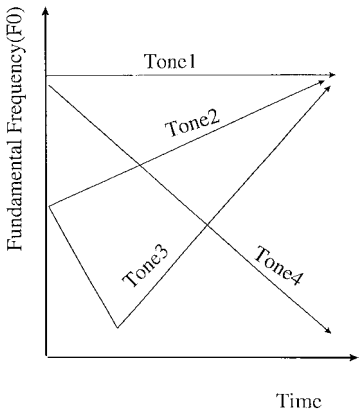

Fig. 2. Standard patterns of the F0 contours of the first four tones.

information for training TTS prosody synthesizers [19], [48], [53]. The other model is designed to predict the prosodic phrase structure embedded in text by using linguistic features extracted from the text [42]–[44], [54], [55]. Obviously, its main use in TTS is to help generate prosodic information.

In the past few years, many studies have been published on deriving prosodic models of spoken language for TTS [19], [30], [31], [44], [47], [53]. Ostendorf and Veilleux [44] used a hierarchical stochastic model to automatically predict prosodic phrasal boundaries in text, achieving promising results in determining where major and minor prosodic breaks occur in input text. Sanders and Taylor [47] identified phrasal breaks in text using a statistical model that described the relationship between phrase breaks and part-of-speech (POS) trigrams. Although these two methods are potentially suitable for use in TTS synthesis, further studies on assigning proper prosodic parameter patterns to the detected prosodic phrases are still needed. Mixdorff and Fujisaki [19], [53] studied an approach based on F0 generation. Their method first locates prosodic phrases in input texts using a syntactical analysis and then applies rules for assigning accent and phrase commands to generate the F0 contour. In [30], an automatic data-driven approach to prosodic modeling was proposed. It automatically explores the relationship between syllabic prosodic patterns and syllable-independent coefficients from a large speech corpus in order to generate proper syllabic prosodic patterns. In [31], a method for modeling the contextual effect of dialog prosody was proposed. It uses linear regression to derive rules for modifying the sentential F0 contours generated by conventional methods for individual sentences.

This general problem of lack of an appropriate prosodic model was encountered in Mandarin TTS prosodic information synthesis. Mandarin Chinese is a tonal language. Each character is pronounced as a syllable. Only about 1300 phonetically distinguishable syllables comprise the set of all legal combinations of 411 base-syllables and five tones. Each base-syllable is composed of an optional consonant *initial* and a vowel *final*. The word, which is the smallest syntactically meaningful unit, consists of one to several syllables. Because syllables are the basic pronunciation units in Mandarin speech, they are also commonly chosen as the basic synthesis units in Mandarin TTS systems. Accordingly, the prosodic

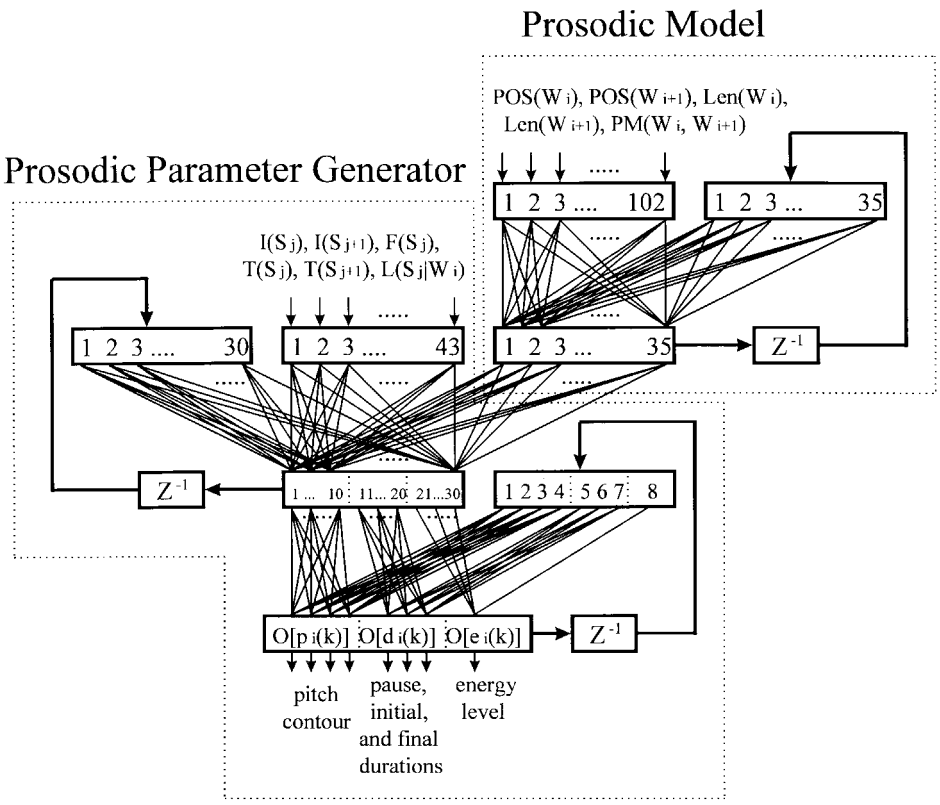

Fig. 3. Detailed architecture of the proposed RNN shown in Fig. 1.

information that must be synthesized includes syllable pitch (or F0) contour, syllable energy contour, syllable *initial* and *final* durations, as well as intersyllable pause duration. Among them, syllable pitch contour has the most important effect on naturalness of synthetic speech. So pitch contour synthesis is of primary concern in Mandarin TTS. Due to its importance, we now briefly discuss the properties of syllable pitch contours in continuous Mandarin speech.

It is known that the tone of a syllable is mainly determined by its pitch contour. Previous studies [56] have concluded that the F0 contour of each of the first four tones can be simply represented by the standard pattern shown in Fig. 2. As for the fifth tone, pronunciation is usually highly context-dependent, so that its F0 contour shape is relatively arbitrary. Nevertheless, it is always pronounced short and light. It would therefore seem that syllable pitch contours in continuous speech are pronounced more consistently so as to make their syntheses in Mandarin TTS systems much simpler. However, in practice, syllable pitch contours are subject to various modifications in continuous speech. So, pitch contour synthesis is not a trivial task. In the past, several methods [9], [10], [57]–[64] have been proposed to synthesize some or all of these prosodic parameters. They include rule-based methods [9], [10], [57]–[59], [63], [64], statistical model-based methods [60], and MLP-based methods [61], [62]. Although these methods have made advances, they are still far away from reaching the goal of generating proper prosodic information for synthesizing natural-sounding speech reproduction of input Chinese text. Their main drawback still lies in their inability to elegantly invoke higher-level linguistic features in exploring the prosodic phrase structure of Mandarin speech to assist in prosodic information generation. This motivate us to construct a more sophisticated prosodic model in this study for developing a high performance Mandarin TTS system.

III. The Proposed RNN-Based Prosodic Information Synthesizer

A multilayer RNN was used to implement the model of the human prosody pronunciation mechanism. The block diagram of the RNN has been depicted in Fig. 1. Its detailed architecture is shown in Fig. 3. As Fig. 3 shows, the RNN is a four-layer network with one input layer, two hidden layers, and one output layer. It can be functionally divided into two parts. The first part consists of a portion of the input layer and the first hidden layer with all outputs being fed back as inputs to itself. It may be considered a prosodic model for exploring the prosodic phrase structure of spoken Mandarin Chinese using only word-level linguistic features of input texts. It operates with word-synchronized clock to generate outputs representing current-word phonologic states of prosodic phrase structures. Input features include POS's POS(W_i) and POS(W_{i+1}), and lengths $Len(W_i)$ and $Len(W_{i+1})$, of both the current word W_i and the following word W_{i+1}, and an indicator, PM(W_i, W_{i+1}), showing the type of punctuation mark (PM) located after the current word. In this work, 42 POS types [65], [66] and four PM types are used. They are listed in Tables I and II, respectively. As Table I shows, the POS set consists of 15 types of verb, eight types of noun, ten types of adverb, two types of conjunction, and seven other

TABLE I
42 POS TYPES USED IN THIS STUDY

1.	Active Intransitive Verb(VA)
2.	Active Pseudo-Transitive Verb(VB)
3.	Active Transitive Verb(VC)
4.	Ditransitive Verb(VD)
5.	Active Verb with a Sentential Object(VE)
6.	Active Verb with a Verbal Object(VF)
7.	Classificatory Verb(VG)
8.	Stative Intransitive(VH)
9.	Sative Pseudo-Transitive Verb(VI)
10.	Stative Transitive Verb(VJ)
11.	Stative Verb with a Sentential Object(VK)
12.	Stative Verb with a Verbal Object(VL)
13.	Nonpredicative Adjective(A)
14.	General Noun(NA)
15.	Special Noun(NB)
16.	Place Noun(NC)
17.	Time Noun(ND)
18.	Determiner(NE)
19.	Measure(NF)
20.	Localizer(NG)
21.	Pronoun(NH)
22.	Adverb of Quantity(DA)
23.	Adverb of Evaluation(DB)
24.	Negation(DC)
25.	Adverb of Time(DD)
26.	Adverb of Degree(DE)
27.	Adverb of Place(DF)
28.	Adverb of Manner(DG)
29.	Aspectual Adverb(DI)
30.	Interrogative Adverb(DJ)
31.	Sentential Adverb(DK)
32.	Preposition(P)
33.	Coordinate Conjunction(CA)
34.	Correlative Conjunction(CB)
35.	Particle(T)
36.	Interjection(I)
37.	Bound(B)
38.	Verb-Complement Compound(VR)
39.	Sentence
40.	Special Verb1(is,are,am)(V1)
41.	Special Verb2(has,have)(V2)
42.	Determiner-Measure Compound(DM)

TABLE II
FOUR GENERAL TYPES OF PUNCTUATION MARK

1	. ;	2	,	3	: !	4	?

POS types. It is noted that the POS set used in this study is a subset of the complete POS set described in [66].

The second part of the RNN consists of the other part of the input layer, the second hidden layer, and the output layer. It is the real prosodic parameter generator. It operates on a syllable-synchronized clock to generate all prosodic parameters that

TABLE III
SIX GENERAL TYPES OF CONSONANT *INITIAL*

1	m, n, l, r, "null"	4	ji, j, tz
2	h, shi, sh	5	p, t, k
3	b, d, g	6	chi, ch, ts, f, s

TABLE IV
SEVENTEEN GENERAL TYPES OF VOWEL *FINAL*

1	a, ia, ua	10	ang, iang, uang
2	o, uo	11	eng, ing, ueng, iong
3	e, ie, iue	12	i
4	ai, iai, uai	13	u
5	ei, uei	14	iu
6	au, iau	15	el
7	ou, iou	16	(ng1)
8	an, ian, uan, iuan	17	(ng2)
9	en, in, uen, iun		

are needed by our Mandarin TTS system, using syllable-level linguistic features fed directly into the second hidden layer as additional inputs, along with outputs from the first part. All outputs of the second hidden layer are fed back as inputs to itself. The output prosodic parameters are also fed back as the inputs to the output layer. This arrangement makes the prosodic parameter generator a dynamic system able to predict time-varying prosodic parameters of real speech. Note also that to reduce the system complexity, nodes in both the output layer and the second hidden layer are partitioned into three groups according to the properties of the eight output prosodic parameters. Output nodes in these three groups correspond to the four parameters of pitch contour, one parameter representing energy level, and three durational parameters, respectively. Input syllable-level linguistic features used in this study include the tone $T(S_j)$, the *initial* type $I(S_j)$, and the *final* type $F(S_j)$ of the current syllable S_j; the tone $T(S_{j+1})$ and the *initial* type $I(S_{j+1})$ of the following syllable S_{j+1}, and an indicator, $L(S_j|W_i)$, showing whether the current syllable forms a monosyllabic word or is the first, an intermediate, or the last syllable of a polysyllabic word. In this study, six broad types of *initial* dependent upon the manner of consonant articulation and 17 types of *final* classified according to the constituent vowel nucleus and nasal ending were used. Tables III and IV list these *initial* and *final* types.

The RNN generated a total of eigt output prosodic parameters. They include for the current syllable: four parameters representing the pitch contour, one parameter representing the energy level (i.e., maximum log-energy), and two parameters representing, respectively, the *initial* and *final* durations; preceding the current syllable: one parameter representing the pause duration. Using four parameters to represent the pitch contour of a syllable is based on results obtained in other studies [60], [62], [67], [68]. We now briefly discuss the pitch-contour parameterization method. As mentioned previously, there are only five basic tones in Mandarin Chinese. The tonality of a syllable is characterized mainly by its pitch contour. Although syllable pitch contours in continuous speech are subject to various modifications, they are all smooth curves with shapes for the first four tones roughly matching corre-

sponding standard tone patterns. We can therefore consider the pitch contour of each syllable as a pattern represented by certain parameters. Specifically, the pitch contour of a syllable is represented by a smooth curve formed through orthonormal polynomial expansion using coefficients up to the third order. The zeroth-order coefficient represents the mean of the pitch contour and the other three coefficients represent its shape. The basis functions of the orthonormal polynomial expansion are expressed as [67]

$$\Phi_0\left(\frac{i}{N}\right)=1 \tag{1}$$

$$\Phi_1\left(\frac{i}{N}\right)=\left[\frac{12\cdot N}{(N+2)}\right]^{1/2}\cdot\left[\left(\frac{i}{N}\right)-\frac{1}{2}\right] \tag{2}$$

$$\Phi_2\left(\frac{i}{N}\right)=\left[\frac{180\cdot N^3}{(N-1)(N+2)(N+3)}\right]^{1/2}\cdot\left[\left(\frac{i}{N}\right)^2-\left(\frac{i}{N}\right)+\frac{N-1}{6\cdot N}\right] \tag{3}$$

$$\Phi_3\left(\frac{i}{N}\right)=\left[\frac{2800}{(N-1)(N-2)(N+2)}\right]^{1/2}\cdot\left[\frac{N^5}{(N+3)(N+4)}\right]^{1/2}\cdot\left[\left(\frac{i}{N}\right)^3-\frac{3}{2}\left(\frac{i}{N}\right)^2+\frac{6N^2-3N+2}{10\cdot N^2}\left(\frac{i}{N}\right)-\frac{(N-1)(N-2)}{20\cdot N^2}\right] \tag{4}$$

for $0\le i\le N$, where $N+1$ is the length of the pitch contour and $N\ge 3$. These basis functions are, in fact, discrete Legendre polynomials. The pitch contour, $Pitch_k(i)$, $0\le i\le N$, of the kth syllable can thus be approximated by

$$Pitch_k'(i)=\sum_{j=0}^{3}p_j(k)\cdot\Phi_j\left(\frac{i}{N}\right) \tag{5}$$

for $0\le i\le N$, where

$$p_j(k)=\frac{1}{N+1}\sum_{i=0}^{N}\Phi_j\left(\frac{i}{N}\right)\cdot Pitch_k(i) \tag{6}$$

is the jth-order coefficient.

Note that all eight output prosodic parameters are further normalized in order to reduce the system complexity resulting from variations in these prosodic parameters caused by lexical phonetic features. This may make training easier. In this study, both the energy level and the *final* duration are normalized for the current syllable *final* type. The *initial* duration and the intersyllable pause duration are normalized for the current syllable *initial* type. The pitch contour is normalized for the current syllable tone type.

The RNN prosody synthesizer can be trained using EBP algorithm [6] on a large set of real-speech utterances. Due to the fact that it is generally difficult to determine analytically the number of hidden nodes of a neural network, the numbers of hidden nodes in both the first and second hidden layers of the RNN were empirically decided. All hidden nodes use sigmoid activation functions. The output layer consists of eight output nodes, all with linear activation functions, to generate the eight prosodic parameters. Because the three types of prosodic information have different dynamic ranges, a distortion measure taken as the objective function for minimization is defined as

$$J(k)=\sum_{j=0}^{3}\{T[p_j(k)]-O[p_j(k)]\}^2+\{T[e(k)]-O[e(k)]\}^2+\sum_{j=0}^{2}\{T[d_j(k)]-O[d_j(k)]\}^2 \tag{7}$$

where $O[p_j(k)]$, $0\le j\le 3$, $O[e(k)]$, $O[d_0(k)]$, $O[d_1(k)]$, and $O[d_2(k)]$ are the outputs of the prosodic information synthesizer (see Fig. 3); and $T[p_j(k)]$, $0\le j\le 3$, $T[e(k)]$, $T[d_0(k)]$, $T[d_1(k)]$, and $T[d_2(k)]$ are the corresponding desired target values which are the normalized parameters of $p_j(k)$, $0\le j\le 3$, $e(k)$, $d_0(k)$, $d_1(k)$, and $d_2(k)$ representing, respectively, the pitch contour, energy level, *initial* duration, *final* duration, and preceding pause duration of the kth syllable. Normalizations of these parameters are defined by

$$T[p_j(k)]=[p_j(k)-m_{p_j}^{t(k)}]/\sigma_p^{t(k)},\qquad 0\le j\le 3, \tag{8}$$

$$T[e(k)]=[e(k)-m_e^{f(k)}]/\sigma_e^{f(k)} \tag{9}$$

$$T[d_j(k)]=[d_j(k)-m_{d_j}^{i(k)}]/\left[\sqrt{3}\sigma_{d_j}^{i(k)}\right]\qquad j=0\text{ and }2 \tag{10}$$

and

$$T[d_1(k)]=[d_1(k)-m_{d_1}^{f(k)}]/\left[\sqrt{3}\sigma_{d_1}^{f(k)}\right] \tag{11}$$

where $i(k)$, $f(k)$, and $t(k)$ are the *initial* type, *final* type, and tone type of the kth syllable, respectively. Here, m_x^l and $(\sigma_x^l)^2$ are the mean and variance of the parameter x, and they are given by

$$m_{p_j}^l=\frac{1}{K_l}\sum_{\substack{k=1\\ t(k)=l}}^{K_l}p_j(k) \tag{12}$$

and

$$(\sigma_p^l)^2=\sum_{j=0}^{3}\frac{1}{K_l}\sum_{\substack{k=1\\ t(k)=l}}^{K_l}[p_j(k)-m_{p_j}^l]^2 \tag{13}$$

for the pitch contour for syllables belonging to the lth tone type;

$$m_e^l=\frac{1}{K_l}\sum_{\substack{k=1\\ f(k)=l}}^{K_l}e(k) \tag{14}$$

and

$$(\sigma_e^l)^2=\frac{1}{K_l}\sum_{\substack{k=1\\ f(k)=l}}^{K_l}[e(k)-m_e^l]^2 \tag{15}$$

for the energy level for syllables belonging to the lth *final* type;

$$m_{d_j}^l = \frac{1}{K_l} \sum_{\substack{k=1 \\ i(k)=l}}^{K_l} d_j(k) \tag{16}$$

and

$$(\sigma_{d_j}^l)^2 = \frac{1}{K_l} \sum_{\substack{k=1 \\ i(k)=l}}^{K_l} [d_j(k) - m_{d_j}^l]^2 \tag{17}$$

$j = 0$ and 2, for the *initial* duration and the preceding pause duration for syllables belonging to the lth *initial* type; and

$$m_{d_1}^l = \frac{1}{K_l} \sum_{\substack{k=1 \\ f(k)=l}}^{K_l} d_1(k) \tag{18}$$

and

$$(\sigma_{d_1}^l)^2 = \frac{1}{K_l} \sum_{\substack{k=1 \\ f(k)=l}}^{K_l} [d_1(k) - m_{d_1}^l]^2 \tag{19}$$

for the *final* duration for syllables belonging to the lth *final* type. It is noted that the scaling factor of $1/\sqrt{3}$ in (10) and (11) is used to make certain the three output prosodic parameter groups have approximately equal contributions to the objective function of the EBP training algorithm.

With normalization, the variations in these prosodic parameters caused by local phonetic structures of individual Mandarin syllables can be greatly reduced. This makes training easier. By feeding-in the linguistic features extracted from the input text as inputs, and setting the normalized prosodic parameters extracted from the corresponding training utterances as the desired output targets, the RNN can be trained to automatically learn and retain the relationships between the prosodic parameter sequences of the training utterances and the linguistic feature sequences of associated texts. A well-trained RNN can therefore be used as a prosody synthesizer for generating proper prosodic parameters for given input texts. Of course, denormalizations of the outputs of the prosodic information synthesizer must be performed in the synthesis process.

IV. Simulations

Performance of the new method of prosodic information synthesis for Mandarin TTS systems was examined through simulations. A continuous-speech Mandarin database provided by the Telecommunication Laboratories, MOTC,[1] R.O.C. was used. The data base consists of four sets of utterances. The first one contains 112 phonetically balanced short sentential utterances with lengths less than 13 syllables. The second set comprises 315 specially designed short utterances with lengths less than 40 syllables. The third and fourth sets comprises, respectively, 28 short and 200 long paragraphic utterances whose texts are all news selected from a large news corpus to cover a variety of subjects including business (12.5%), medicine (12%), social events (12%), sports (10.5%), literature (9%), computers (8%), food and nutrition (8%), movies (6.5%), family life (6.5%), tours (6%), politics (2.5%), traffic and transportation (2.5%), etc. All utterances were generated by a single male speaker. They were all spoken naturally at a speed of 3.5 to 4.5 syllables/s. The data base was divided into two parts: a training set and an open test set. These two sets consisted of 28 191 and 7051 syllables, respectively.

TABLE V
RMSE's of the Five Types of Synthesized Prosodic Information

	Close Test	Open Test
Pitch Contour	0.84ms/Frame	1.06ms/Frame
Energy Level	3.39dB	4.17dB
Initial Duration	17.2ms	18.5ms
Final Duration	33.3ms	36.7ms
Pause Duration	23.7ms	54.5ms

All speech signals were digitally recorded using a 20-kHz sampling rate. They were then divided into 10-ms frames and manually segmented into silence, unvoiced, and voiced parts according to observation of acoustic features including waveforms, energy, zero crossing rates, LPC coefficients, cepstra and delta-cepstra. The eight prosodic parameters to be synthesized for each syllable were then extracted from the downsampled 10-kHz speech signals. They included the four orthogonally transformed coefficients of pitch contour, maximal log-energy, *initial* duration, *final* duration, and preceding pause duration. Here, pitch period was detected using the SIFT algorithm [69] with manual error-correction. The frame length for pitch detection was 40-ms with a 10-ms frame shift. The frame length for log-energy analysis was 20-ms with a 10-ms frame shift. Both cases used rectangular windows.

An automatic tagging algorithm based on the criterion of long-word-first was then used to segment all the texts associated with the training utterances in the speech data base to obtain the word sequences. A Chinese lexicon containing approximately 80 000 words[2] was used in the tagging. Words in the lexicon consist of one to five syllables. All tagging errors were manually corrected. The POS's of all words were then manually determined. As mentioned before, the set of 42 POS types listed in Table I was used in this study. Finally, all linguistic features were extracted for use in the system.

The RNN prosody synthesizer was trained using the EBP algorithm. The numbers of nodes in the first and second hidden layers were determined empirically and set to be 35 and 30, respectively. The learning rates for training the two types of weights connecting to hidden nodes and to output nodes were initially set to be 0.01 and 0.001, respectively. They were all linearly decayed to zero at 200 training epochs. The training process converged approximately after 50 training epochs. It took about 10 h run on a DEC 3000 workstation.

Table V lists the root mean square errors (RMSE's) of the synthesized prosodic parameters. It shows that RMSE's of 0.84 and 1.06 ms/frame were achieved in pitch contour synthesis for the closed and the open tests, respectively. A typical example of pitch mean synthesis for the open test is shown in Fig. 4(a). It can be seen from the figure that the trajectories of the

[1] The Ministry of Transportation and Communications

[2] The lexicon was supplied by the Institute of Information Science, Academia Sinica.

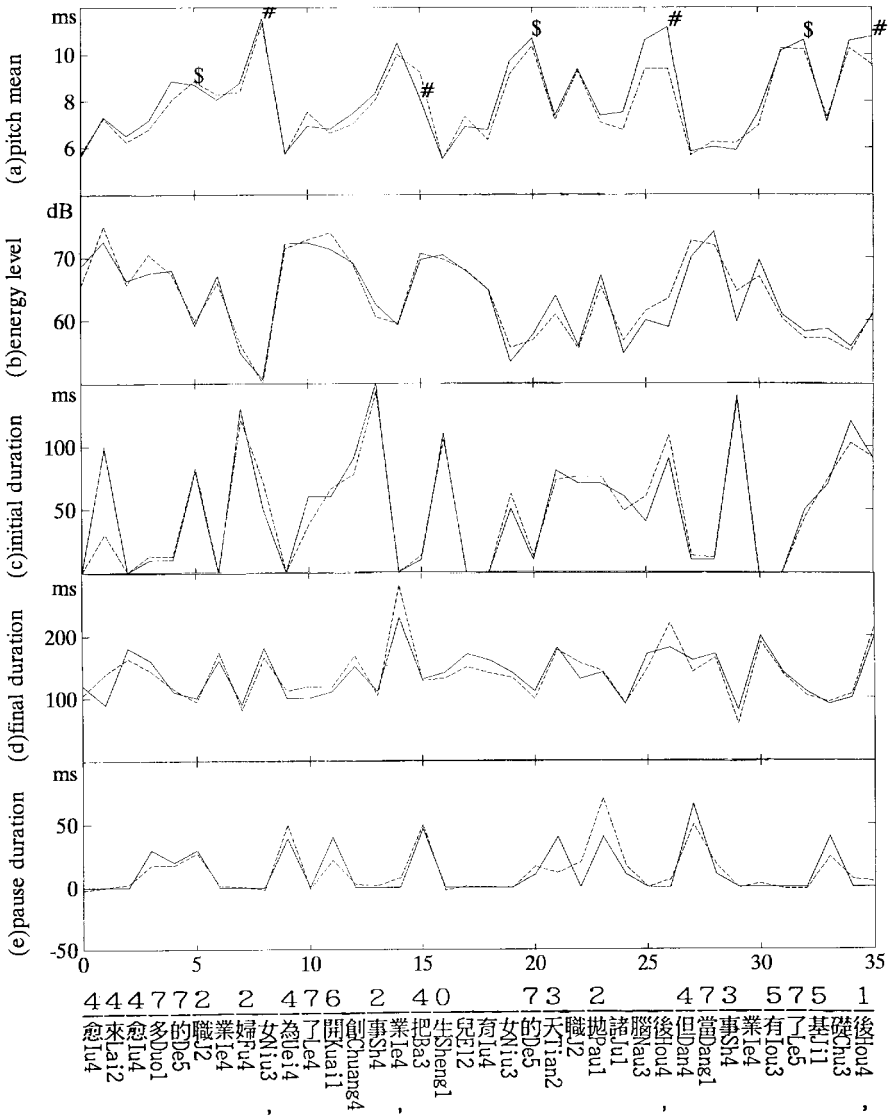

Fig. 4. Typical example of the original (solid lines) and the synthesized (dotted lines) prosodic parameter sequences of: (a) the pitch mean, (b) energy level, (c) *initial* duration, (d) *final* duration of syllable; and (e) intersyllable pause duration. The text is: "More and more women are concentrating on career development and delaying fulfillment of their natural childbearing and child-rearing functions. After they have achieved their career goals however, they then find their ovaries have deteriorated. This is creating an urgent demand for viable ova, but ovum donation, unlike sperm donation, can be painful and dangerous since ova are taken through the abdominal wall, or via ultrasound techniques, and there is always a risk of adverse reaction to anesthesia." Note that the x-axis represents the syllable sequence, and the broken line and the numbers on the top of the text show, respectively, the segmentation of the syllable sequence into a word sequence and the associated state sequence.

synthesized pitch means match quite well with their original counterparts for most syllables. Through further error analysis, we found that only few large errors had occurred in the pitch mean synthesis. Most of them take place at last syllables of sentences of Tone 3, and result mainly from extraordinary Tone 3 pronunciations which generate extremely large pitch means. Some other large errors occur at syllables with Tone 5. Because most of them are caused by alternative but legal Tone 5 pronunciations, they are not serious. In energy level synthesis, RMSE's of 3.39 and 4.17 dB were obtained for the closed and the open tests, respectively. Fig. 4(b) shows the energy-level-synthesis results for the input text used in Fig. 4(a). Clearly, the trajectories of the synthesized energy levels also match quite well with their original counterparts for most syllables. In *initial*-duration synthesis, RMSE's of 17.2 and 18.5 ms were obtained for the closed and the open tests, respectively. Fig. 4(c) shows the synthesized syllable *initial* durations for the same input text used previously. In the figure we see that the trajectories of synthesized *initial* durations also match very well with their original counterparts for most syllables. In *final*-duration synthesis, RMSE's of 33.3 and 36.7 ms were obtained for the closed and the open tests, respectively. Fig. 4(d) shows the synthesized syllable *final* durations for the same input text used previously. Again we find in the figure that the trajectories of the synthesized *final* durations match very well with their original counterparts for most syllables. In intersyllable pause duration synthesis, both the training and the test processes were slightly modified for cases in which a PM existed. During training, the term $\{T[d_2(k)] - O[d_2(k)]\}^2$ in (7) was simply set to zero for this special case in order

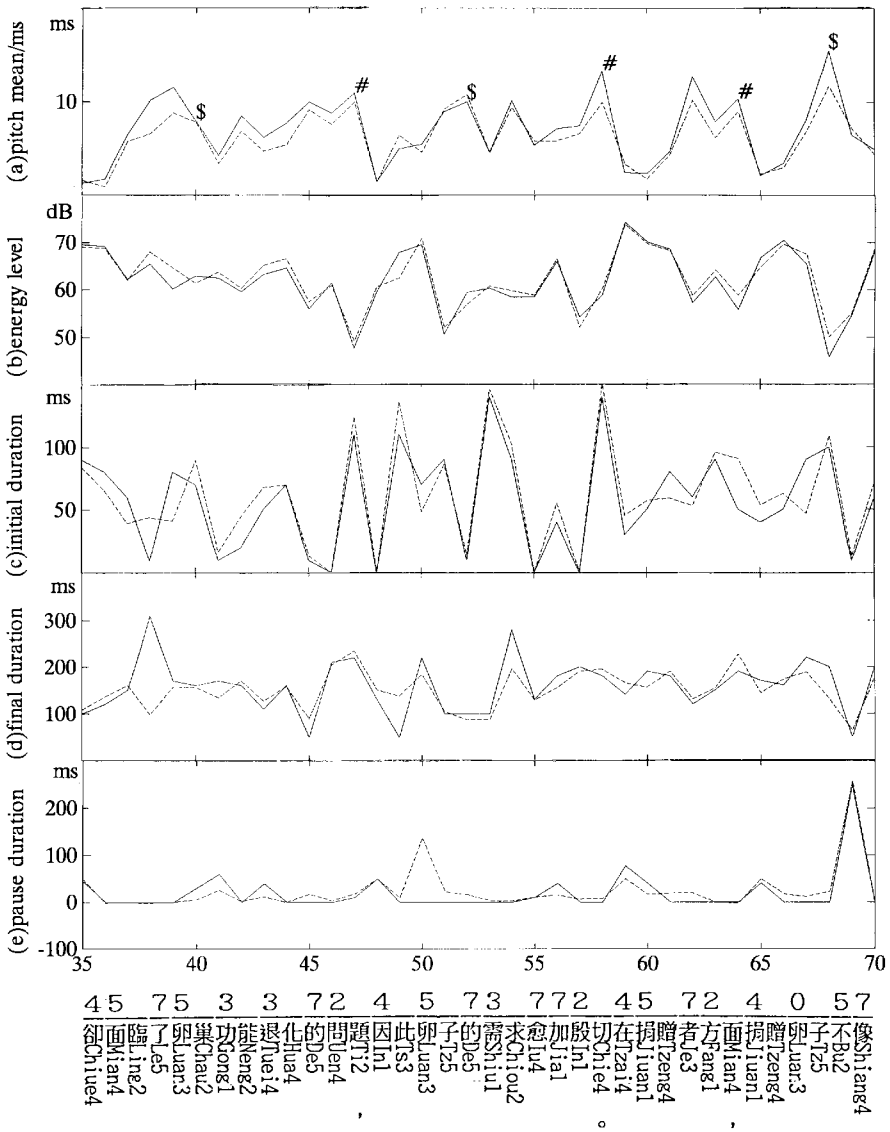

Fig. 4. (*Continued*).

to eliminate its dominant effect on the weight adjustment. During testing, a constant pause duration ($= 50$ ms) was used whenever a PM occurred. Experimental results are also shown in Table V. RMSE's of 23.7 and 54.5 ms were obtained for the closed and the open tests, respectively. Fig. 4(e) shows the synthesized intersyllable pause durations for the same input text used previously. In the figure we see that the trajectories of the synthesized pause durations match reasonably well with their original counterparts for most syllables. Many major breaks were correctly set at proper locations without any PM. Although a few large mismatch errors occurred, some of the mismatching long breaks synthesized may still properly represent alternative breathing breaks. An informal listening test (discussed below) confirmed that only very few long breaks were unnaturally set.

Two typical examples of the pitch contour synthesis in the open test are shown in Fig. 5(a) and (b), respectively. We see in these figures that most synthesized syllable pitch contours resemble their original counterparts in both shape and level. It is worth noting that in Fig. 5(a), the synthesized pitch contour shapes of syllables 4 to 7, whose lexical tonalities are all Tone 3, look like the standard patterns of Tone 2, Tone 3, Tone 2, and Tone 3, respectively. Actually, these four syllables form two bisyllabic words. Similarly, in Fig. 5(b), the synthesized pitch contour shapes of the last two syllables, whose lexical tonalities are all Tone 3, look like the standard patterns of Tone 2 and Tone 3, respectively. But here these two syllables belong to different words. These results show that the famous Sandhi rule of changing a Tone 3 to a Tone 2 when it is followed by a Tone 3 has been correctly implemented here. By careful listening of both the original and synthesized pitch contours of all syllable sequences with 3-3 tone pair and 3-3-3 tone sequence in the data base, we were able to relabel the tonalities of all Tone 3 syllables with the actual tones being pronounced, and to calculate the number of tone changes. Table VI lists the experimental results. As shown in the table, the rate of correct synthesis was 86% for 3-3 tone pairs and 77.4% for 3-3-3 tone sequences. Through further error analysis, we found that most errors occurred at syllables that can be pronounced as either Tone 2 or Tone

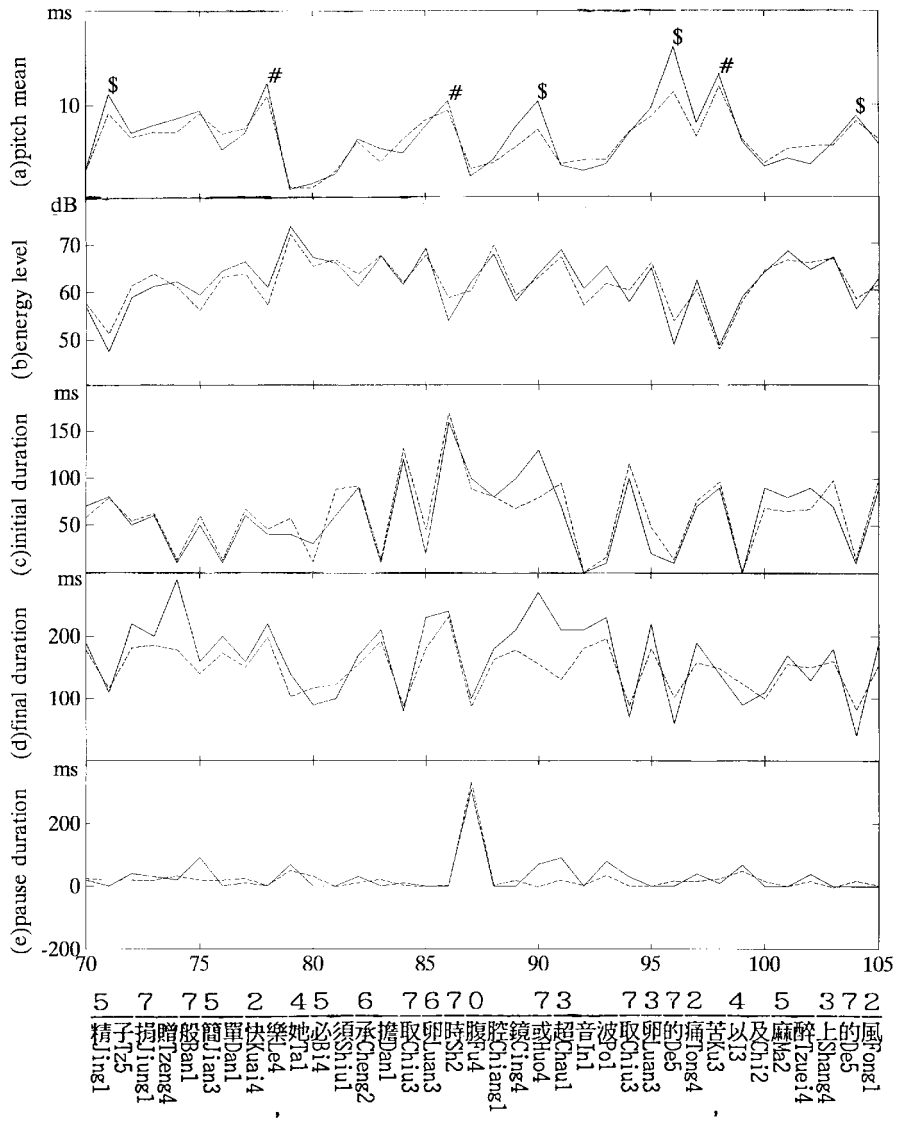

Fig. 4. (*Continued*).

3. So, only very few annoying mispronunciations of Tone 3 were perceived. This confirms that the Sandhi rule for Tone 3 change was automatically learned and implicitly retained by the RNN prosody synthesizer. We then examined some other pitch contour modification phonologic rules. The four tone concatenation rules of $4' - 4, 4 - 3', 3 - 1'$, and $4 - 1'$ proposed in [9] were examined. Here, the superscript "′" denotes that pitch contour mean or shape should be adjusted in order to connect the two pitch contours more smoothly. Fig. 6 shows the average synthesized pitch-contour-pair patterns for these four tone-pairs. We note here that the average pitch-contour-pair pattern of each tone-pair was produced through two orthogonal polynomial expansions using the two sets of four average pitch contour parameters of the tone-pair. It can be seen in the figure that all average patterns for these four synthesized pitch-contour pairs match reasonably well with those of the actual speech. However, as Fig. 6 shows, these four rules are not completely correct and must be modified. This judgement is based on comparing the average pattern of each pitch-contour pair in the actual speech with that formed by directly concatenating the average pitch contours of the corresponding two tones. Specifically, the three tone concatenation rules of $4' - 4, 4 - 3'$, and $4 - 1'$ should be changed to $4' - 4'$, $4' - 3'$, and $4' - 1'$, respectively. Last, the synthesis rules for Tone 5 were examined. The most significant property of Tone 5 pronunciation is the relatively low energy level and short duration. By examining the average values of the synthesized energy levels and syllable *final* durations for all tones, we found that Tone 5 has the lowest energy level and the shortest *final* duration. These results match our linguistic knowledge. Based on above discussion, we can therefore conclude that many human F0 phonologic rules were learned by the proposed RNN prosody synthesizer.

To examine the characteristics of the prosodic model in greater detail, we vector-quantized the outputs of the first part of the RNN prosody synthesizer into eight classes and assigned each class a state, forming an eight-state finite state automaton (FSA). This FSA operates according to a clock synchronized with input text words. Table VII lists some FSA statistics, including the state transition probabilities, the distributions

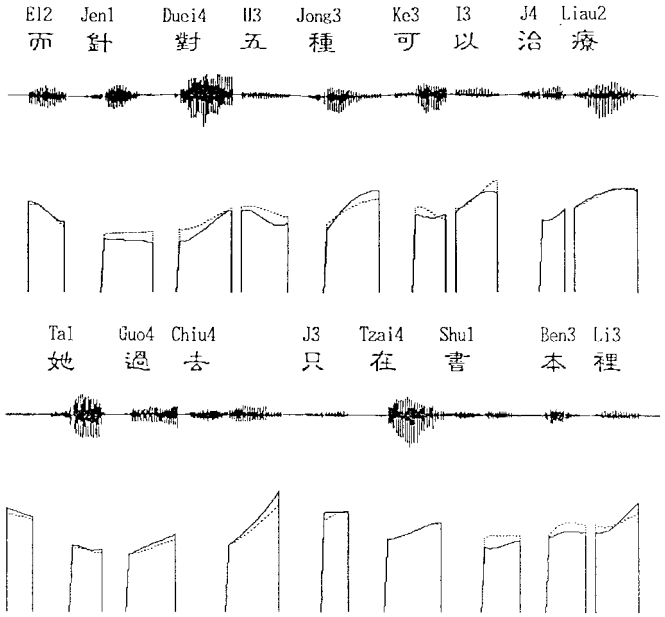

Fig. 5. Two typical synthesized syllable pitch contour sequences.

TABLE VI
(a) TONE-3–CHANGE STATISTICS IN THE SYNTHESIZED PITCH CONTOURS OF SYLLABLE SEQUENCES WITH 3–3 TONE PAIRS. (b) TONE-3–CHANGE STATISTICS IN THE SYNTHESIZED PITCH CONTOURS OF SYLLABLE SEQUENCES WITH 3–3–3 TONE SEQUENCES

Pronounced Tone Pair	Tone 2-3			
Synthesized Tone Pair	Tone 2-3		Tone 3-3	
Condition of the Pause Between the Two Tones	Intra Word	Inter Word	Intra Word	Inter Word
Number	330	199	24	48
Probability	0.49	0.29	0.04	0.07
Pronounced Tone Pair	Tone 3-3			
Synthesized Tone Pair	Tone 2-3		Tone 3-3	
Condition of the Pause Between the Two Tones	Intra Word	Inter Word	Intra Word	Inter Word
Number	5	20	1	53
Probability	0.01	0.03	0.00	0.08

(a)

Pronounced Tone Sequence	Tone 2-2-3			
Synthesized Tone Sequence	Tone 2-2-3		Tone 3-2-3	
Condition of the Pause Between the First two Tones	Intra Word	Inter Word	Intra Word	Inter Word
Number	36	6	4	16
Probability	0.39	0.07	0.04	0.17
Pronounced Tone Sequence	Tone 3-2-3			
Synthesized Tone Sequence	Tone 2-2-3		Tone 3-2-3	
Condition of the Pause Between the First two Tones	Intra Word	Inter Word	Intra Word	Inter Word
Number	1	1	1	28
Probability	0.01	0.01	0.01	0.30

(b)

of beginning and ending words in the sentential and the paragraphic texts, the distributions of words before and after PM's, and the distributions of words of different lengths. Fig. 7 depicts the topology of the FSA as only some most significant state transitions are drawn. Tables VII(b) and VII(c) show that State 1 and State 2 are the ending states of sentences

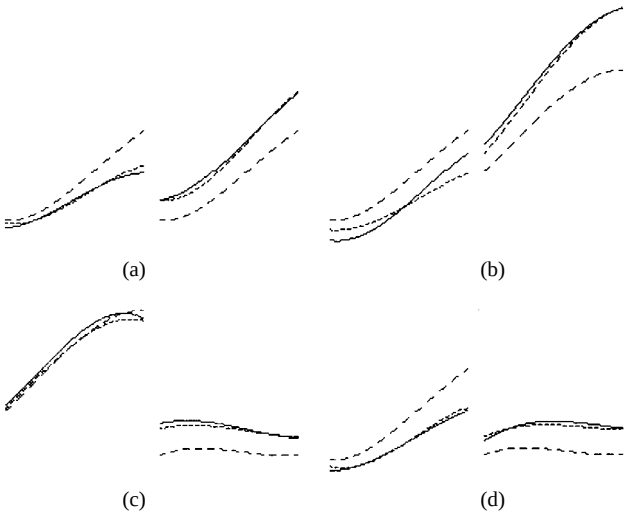

Fig. 6. The average patterns of the original (solid lines) and the synthesized (dotted lines) pitch-contours of 4 tone-pairs: (a) $4'$-4, (b) 4-$3'$, (c) 3-$1'$, and (d) 4-$1'$. The patterns shown by the broken lines are formed by directly concatenating the two average pitch contours of the corresponding tones.

TABLE VII
(a) STATISTICS FROM THE FSA DERIVED BY THE FIRST PART OF THE RNN: STATE TRANSITION PROBABILITIES. (b) STATISTICS FROM THE FSA DERIVED BY THE FIRST PART OF THE RNN: BEGINNING AND ENDING WORDS OF SENTENTIAL AND PARAGRAPHIC TEXTS. (c) STATISTICS FROM THE FSA DERIVED BY THE FIRST PART OF THE RNN: DISTRIBUTIONS OF WORDS BEFORE AND AFTER PM. (d) STATISTICS FROM THE FSA DERIVED BY THE FIRST PART OF THE RNN: DISTRIBUTIONS OF WORDS WITH DIFFERENT LENGTHS

	State0	State1	State2	State3	State4	State5	State6	State7
State0	0.13	0.05	0.02	0.05	0.14	0.16	0.10	0.36
State1	0.06	0.03	0.00	0.03	0.76	0.05	0.02	0.05
State2	0.07	0.01	0.00	0.01	0.75	0.03	0.06	0.04
State3	0.04	0.09	0.04	0.08	0.02	0.07	0.19	0.46
State4	0.10	0.02	0.03	0.05	0.02	0.47	0.05	0.26
State5	0.09	0.06	0.12	0.16	0.00	0.21	0.14	0.21
State6	0.05	0.14	0.20	0.15	0.00	0.06	0.26	0.15
State7	0.07	0.09	0.19	0.18	0.00	0.13	0.21	0.13

(a)

	State0	State1	State2	State3	State4	State5	State6	State7
Begin	0.07	0.04	0.00	0.00	0.85	0.01	0.01	0.02
End	0.03	0.75	0.21	0.01	0.00	0.00	0.00	0.00

(b)

	State0	State1	State2	State3	State4	State5	State6	State7
After PM	0.09	0.06	0.03	0.03	0.74	0.02	0.01	0.02
Before PM	0.14	0.36	0.49	0.01	0.01	0.00	0.00	0.00

(c)

Word Length	State0	State1	State2	State3	State4	State5	State6	State7
1	0.04	0.09	0.00	0.08	0.15	0.12	0.08	0.44
2	0.05	0.06	0.17	0.14	0.15	0.22	0.17	0.05
3	0.31	0.04	0.09	0.07	0.12	0.01	0.33	0.04
4	0.70	0.10	0.12	0.03	0.01	0.01	0.02	0.01
5	0.83	0.00	0.13	0.03	0.00	0.00	0.00	0.00

(d)

or paragraphs. State 4 is the beginning state of sentences. From Table VII(d), we find that State 7 is associated with monosyllabic words. State 0 is associated with polysyllabic words with lengths greater than or equal to 3. State 0 is also associated with most proper nouns. Some trisyllabic words are associated with State 6. From Fig. 7 [or Table VII(a)], it can be found that State 5 and State 7 both have a high probability of following State 4, so they appear very often near the beginnings of sentences. State 4 follows State 1 and

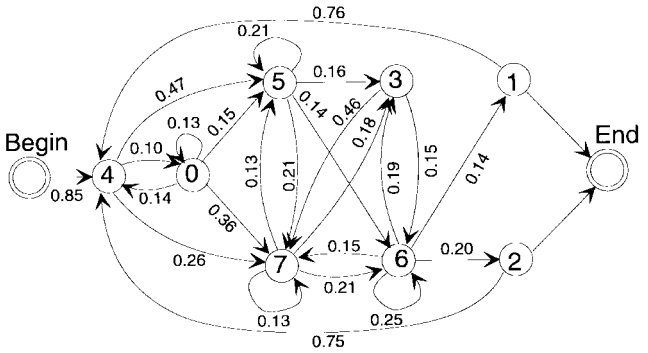

Fig. 7. Topology of the FSA derived from the first part of the proposed RNN prosody synthesizer.

TABLE VIII
(a) STATISTICAL MEANS OF THE THREE PROSODIC PARAMETERS OF PITCH MEAN, ENERGY LEVEL, AND *FINAL* DURATION FOR THE FIRST SYLLABLES OF WORDS IN EACH STATE . (b) STATISTICAL MEANS OF THE THREE PROSODIC PARAMETERS OF PITCH MEAN, ENERGY LEVEL, AND *FINAL* DURATION FOR THE LAST SYLLABLES OF WORDS IN EACH STATE

	State0	State1	State2	State3	State4	State5	State6	State7
Pitch Mean(ms)	7.706	9.525	8.716	8.062	6.651	7.421	7.855	8.242
Energy Level(db)	63.21	60.29	59.80	62.46	66.48	64.43	62.55	62.64
Final Duration(ms)	152.0	192.5	156.2	151.5	122.6	141.6	141.0	133.8

(a)

	State0	State1	State2	State3	State4	State5	State6	State7
Pitch Mean(ms)	8.716	9.972	9.930	8.488	6.866	7.860	8.318	8.316
Energy Level(db)	61.83	59.53	58.43	61.89	67.19	63.78	61.97	62.58
Final Duration(ms)	206.7	230.0	225.1	175.7	150.9	160.7	160.5	137.7

(b)

State 2, usually with a PM located between them. State 7 sometimes follows State 3 forming an adjectival phrase. More interpretations of the FSA may be inducted by observing other texts and their corresponding encoded state sequences as generated by the prosodic model.

We then examined the relationship between the FSA states and the prosodic phrase structure of Mandarin speech. Fig. 4(a) shows the syllable pitch mean trajectory of part of a paragraphic utterance. By manually marking the intonational phrase boundaries with "#" and the intermediate phrase boundaries with "$," we find from the figure that many prosodic phrases are present in the speech segment. By observing the states of the constituent words of these prosodic phrases, we find that an intonational phrase always starts with a State 4 word and ends with a State 1 or State 2 word. Many intermediate phrases start with State 3 words. State 5 and State 7 words often follow State 4 words at the beginnings of intonational phrases. Table VIII(a) and VIII(b) list, respectively, the statistical means of the three prosodic parameters of pitch mean, energy level, and *final* duration for the first and last syllables of words of each state. In Table VIII(a) we see that the first syllable of a State 4 word has, on average, the lowest pitch mean, the highest energy level, and the shortest *final* duration. And the first syllable of a State 5 word has, on average, a lower pitch mean, higher energy level, and shorter *final* duration. By contrast, Table VIII(b) shows that the last syllables of State 1 and State 2 words have, on average, the highest pitch means, the lowest energy levels, and the longest *final* durations. Obviously, these properties conform to the acoustic characteristics of the beginning and ending syllables of intonational phrases. Lastly, by comparing Table VIII(a) and VIII(b), we find that for each state the last syllable of a word always has, on average, a higher pitch mean, lower energy level, and longer *final* duration than first syllables of words of the same state. So, the stress on a word is usually placed on the first syllable.

The above discussion confirms that the FSA is linguistically meaningful. The following two conclusions can therefore be drawn. First, the first part of the RNN, which generates the FSA, is an effective prosodic model for exploring the prosodic phrase structure of Mandarin Chinese. Second, the effects of high-level linguistic features on prosodic information generation must have been properly considered by the proposed prosodic model using only word-level linguistic feature inputs.

Lastly, a pitch-synchronous-overlap-add (PSOLA) based Mandarin TTS system was used for subjective testing of the proposed RNN prosody synthesizer. It was realized in real-time on a PC/AT 486 with a 16-b Sound Blaster add-on card. It used a set of 411 base-syllable waveforms extracted from the training data set as the basic synthesis units. Statistical model-based text analysis was used to automatically tag the input text to generate all linguistic features needed by the system. We note that the POS's of all words were automatically generated by this system. Three groups of prosodic information including pitch contour, energy level, and three durational parameters were generated by the proposed RNN prosodic information synthesizer. Informal listening tests using many long input texts not included in the database, conducted with many native Chinese-speakers living in Taiwan confirmed that all synthesized speech sounded very natural. Based on those tests, we can therefore conclude that the proposed RNN prosody synthesizer performed very well and is practically useful for Mandarin TTS systems.

V. CONCLUSIONS

A new neural network-based prosodic information synthesizer for Mandarin TTS has been discussed in this paper. It employs a compact four-layer RNN that simultaneously generates prosodic information including syllable pitch contour, energy level, *initial* and *final* durations, as well as intersyllable pause duration. Trained on a large set of sentential and paragraphic utterances, the RNN prosody synthesizer learned many human phonologic rules including the well-known Sandhi Tone 3 change rule. Detailed analysis of its hidden layer activities revealed that the well-trained RNN has the ability to track phonologic states of the prosodic phrase structure of the speech being synthesized from input texts. So, the effects of high-level linguistic features on prosodic information generation are well handled by the RNN. Experimental results showed that most synthesized parameter sequences match very well with their original counterparts. A PSOLA-based Mandarin TTS system was also developed for further evaluating RNN performance. Informal listening tests involving many native Chinese-speakers confirmed that

all synthesized speech sounded very natural. So, the proposed RNN is a promising prosodic information synthesizer for Mandarin TTS systems.

ACKNOWLEDGMENT

The authors thank the Telecommunication Laboratories, MOTC, for supplying the speech data base, and Academia Sinica for supplying the lexicon.

REFERENCES

[1] D. Servan-Schreiber, A. Cleermans, and J. L. McClelland, "Graded state machines: The representation of temporal contingencies in simple recurrent networks," *Mach. Learn.*, vol. 7, pp. 161–193, 1991.
[2] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, pp. 179–211, 1990.
[3] ——, "Distributed representations, simple recurrent networks, and grammatical structure," *Mach. Learn.*, vol. 7, pp. 195–224, 1991.
[4] ——, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, pp. 71–99, 1993.
[5] D. E. Rumelbert, G. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing, Vol. 1: Foundations.* Cambridge, MA: MIT Press, 1986.
[6] S. J. Lee, K. C. Kim, H. Yoon, and J. W. Cho, "Application of fully recurrent neural networks for speech recognition," in *Proc. ICASSP*, 1991, pp. 77–80.
[7] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, pp. 137–181, Sept. 1987.
[8] J. P. Olive, "Fundamental frequency rules for the synthesis of simple declarative English sentences," *J. Acoust. Soc. Amer.*, vol. 57, pp. 476–482, 1975.
[9] L. S. Lee, C. Y. Tseng, and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1309–1320, 1989.
[10] 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. Speech Audio Processing*, vol. 1, pp. 287–294, July 1993.
[11] M. Riedi, "A neural-network-based model of segmental duration for speech synthesis," in *Proc. EUROSPEECH*, 1995, pp. 599–602.
[12] Y. Sagisaka, "On the prediction of global F0 shape for Japanese text-to-speech," in *Proc. ICASSP*, 1990, pp. 325–328.
[13] C. Traber, "F0 generation with a database of natural F0 patterns and with a neural network," in *Talking Machines: Theories, Models and Applications.* Amsterdam, The Netherlands: Elsevier, 1992.
[14] M. S. Scordilis and J. N. Gowdy, "Neural network based generation of fundamental frequency contours," in *Proc. ICASSP*, 1989, pp. 219–222.
[15] D. H. Klatt, "The Klattalk text-to-speech conversion system," in *Proc. ICASSP*, 1982, pp. 1589–1592.
[16] J. 't Hart and A. Cohen, "Intonation by rule: A perceptual guest," *J. Phonet.*, vol. 1, pp. 309–327, 1973.
[17] J. P. Olive and H. L. Nakatani, "Rule-synthesis of speech by word concatenation: A first step," *J. Acoust. Soc. Amer.*, vol. 55, pp. 660–666, 1974.
[18] H. Fujisaki, K. Hirose, N. Takahashi, and H. Morikawa, "Acoustic characteristics and the underlying rules of intonation of the common Japanese used by radio and TV announcers," in *Proc. ICASSP*, 1986, pp. 2039–2042.
[19] H. Mixdorff and H. Fujisaki, "A scheme for a model-based synthesis by rule of F0 contours of German utterances," in *Proc. EUROSPEECH*, 1995, pp. 1823–1826.
[20] P. L. Salza, A. Spini, S. Quazza, and M. Falcone, "Prosody of a text-to-speech synthesizer for Italian in a dialogue system model and evaluation," CSELT Tech. Rep., vol. 22, pp. 217–239, Apr. 1994.
[21] G. Epitropakis, N. Yiourgalis, and G. Kokkinakis, "Prosody assignment to TTS-system based on linguistic analysis," in *Proc. 4th Australian Int. Conf. Speech Science and Technology*, 1992, pp. 534–539.
[22] N. Yiourgalis, G. Epitropakis, and G. Kokkinakis, "Some important cues on improving the quality of a TTS system," in *Proc. 4th Australian Int. Conf. on Speech Science and Technology*, 1992, pp. 528–533.
[23] M. H. O'Malley, H. Resnick, and M. Caisse, "An analysis of strategies for finding prosodic clues in text," in *Proc. EUROSPEECH*, 1991, vol. 3, pp. 1165–1168.
[24] B. Horvei, G. Ottesen, and S. Stensby, "Analyzing prosody by means of a double tree structure," in *Proc. EUROSPEECH*, 1993, vol. 3, pp. 1987–1990.
[25] C. Traber, "Syntactic processing and prosody control in the SVOX TTS system for German," in *Proc. EUROSPEECH*, 1993, vol. 3, pp. 2099–2102.
[26] G. Bailly, "Integration of rhythmic and syntactic constraints in a model of generation of French prosody," *Speech Commun.*, vol. 8, pp. 137–146, 1989.
[27] E. Lopez-Gonzalo and L. A. Hernandez-Gomez, "Data-driven joint F0 and duration modeling in text to speech conversion for Spanish," in *Proc. ICASSP*, 1994, vol. 1, pp. 589–592.
[28] F. Emerard *et al.*, "Prosodic processing in a text-to-speech synthesis system using a database and learning procedure," in *Talking Machines: Theories, Models and Applications.* Amsterdam, The Netherlands: Elsevier, 1992.
[29] Y. Yamashita and R. Mizoguchi, "Automatic generation of prosodic rules for speech synthesis," in *Proc. ICASSP*, 1994, vol. 1, pp. 593–596.
[30] E. Lopez-Gonzalo and L. A. Hernandez-Gomez, "Automatic data-driven prosodic modeling for text to speech," in *Proc. EUROSPEECH*, 1995, pp. 585–588.
[31] Y. Yamashita and R. Mizoguchi, "Modeling the contextual effects on prosody in dialog," in *Proc. EUROSPEECH*, 1995, pp. 1329–1332.
[32] N. Kaiki, K. Mimura, and Y. Sagisaka, "Statistical modeling of segmental duration and power control for Japanese," in *Proc. EUROSPEECH*, 1991, pp. 625–628.
[33] T. Fukada, Y. Komori, T. Aso, and Y. Ohora, "A study of pitch pattern generation using HMM-based statistical information," in *Proc. ICSLP*, 1994, pp. 723–726.
[34] S. Fujio, Y. Sagisaka, and N. Higuchi, "Stochastic modeling of pause insertion using context-free grammar," in *Proc. ICASSP*, 1995, pp. 604–607.
[35] T. J. Sejnowski and C. R. Rosenberg, "NETalk: A parallel network that learns to read aloud," Tech. Rep. JHU/EECS-86/01, Johns Hopkins Univ., Baltimore, MD, 1986.
[36] P. Taylor, "Using neural networks to locate pitch accents," in *Proc. EUROSPEECH*, 1995, pp. 1345–1348.
[37] L. F. M. ten Bosch, "Automatic classification of pitch movements via MLP-based estimation of class probabilities," in *Proc. ICASSP*, 1995, pp. 608–611.
[38] A. K. Syrdal, "Text-to-speech systems," in *Applied Speech Technology.* Boca Raton, FL: CRC, 1995.
[39] T. Saito *et al.*, "ProTALKER: A Japanese text-to-speech system for personal computers," IBM TRL Res. Rep. RT0110, June 1995.
[40] Y. Hara, T. Nitta, H. Saito, and K. Kobayashi, "Development of TTS card for PC's and TTS software for WS's," *IEICE Trans. Fund. Elec., Commun., Comput. Sci.*, vol. E76-A, pp. 1999–2007, Nov. 1993.
[41] S. Hunnicutt, "The development of text-to-speech technology for use in communication aids," in *Applied Speech Technology.* Boca Raton, FL: CRC, 1995.
[42] M. Ostendorf, C. W. Wightman, and N. Veilleux, "Parse scoring with prosodic information: On analysis/synthesis approach," *Computer Speech and Language.* New York: Academic, 1993, pp. 193–210.
[43] J. Bachenko and E. Fitzpatrick, "A computational grammar of discourse neutral prosodic parsing in English," *Computat. Linguist.*, vol. 16, pp. 155–170, 1990.
[44] M. Ostendorf and N. Veilleux, "A hierarchical stochastic model for automatic prediction of prosodic boundary location," *Computat. Linguist.*, vol. 20, pp. 27–54, 1994.
[45] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," in *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 469–480, Oct. 1994.
[46] N. Campbell, "Automatic detection of prosodic boundaries in speech," *Speech Commun.*, vol. 13, pp. 343–354, 1993.
[47] E. Sanders and P. Taylor, "Using statistical models to predict phrase boundaries for speech synthesis," in *Proc. EUROSPEECH*, 1995, pp. 1811–1814.
[48] T. Hirai, N. Higuchi, and Y. Sagisaka, "Automatic detection of major phrase boundaries using statistical properties of superpositional F0 control model parameters," in *Proc. EUROSPEECH*, 1995, pp. 1341–1344.
[49] M. Q. Wang and J. Hirshberg, "Automatic classification of intonational phrase boundaries," *Computer Speech and Language.* New York: Academic, 1992, vol. 6, pp. 175–196.
[50] R. Kompe *et al.*, "Prosodic scoring of word hypotheses graphs," in *Proc. EUROSPEECH*, 1995, pp. 1333–1336.
[51] M. Nakai, H. Singer, Y. Sagisaka, and H. Shimodaira, "Automatic prosodic segmentation by F0 clustering using superpositional modeling," in *Proc. ICASSP*, 1995, pp. 624–627.

[52] G. Bruce and B. Granstrom, "Prosodic modeling in Swedish speech synthesis," *Speech Commun.*, vol. 13, pp. 63–73, 1993.
[53] H. Fujisaki and S. Ohno, "Analysis and modeling of fundamental frequency contours of English utterances," in *Proc. EUROSPEECH*, 1995, pp. 985–988.
[54] S. Frenkenberger, B. Schnabel, M. Alissali, and M. Kommenda, "Prosodic parsing based on parsing of minimal syntactic structure," in *Proc. 2nd IEE/ESCA Workshop on Speech Synthesis*, Mohonk, NY, 1994.
[55] R. A. Sharman and J. H. Wright, "A fast stochastic parser for determining phrase boundaries for text-to-speech synthesis," in *Proc. ICASSP*, 1996, pp. 357–360.
[56] Y. R. Chao, *A Grammar of Spoken Chinese.* Berkeley, CA: Univ. California Press, 1968.
[57] J. Zhang, "Acoustic parameters and phonological rules of a text-to-speech system for Chinese," in *Proc. ICASSP*, 1986, pp. 2023–2026.
[58] N. C. Chan and C. Chan, "Prosodic rules for connected Mandarin synthesis," *J. Inform. Sci. Eng.*, vol. 8, pp. 261–281, June 1992.
[59] Y. C. Chang, B. E. Shia, Y. F. Lee, and H. C. Wang, "A study on the prosodic rules for Chinese speech synthesis," in *Proc. Int. Conf. Computer Processing of Chinese and Oriental Languages*, Aug. 1991, pp. 210–215.
[60] S. H. Chen, S. G. Chang, and S. M. Lee, "A statistical model based fundamental frequency synthesizer for Mandarin speech," *J. Acoust. Soc. Amer.*, vol. 92, pp. 114–120, July 1992.
[61] S. H. Hwang and S. H. Chen, "Neural network synthesizer of pause duration for Mandarin text-to-speech," *Electron. Lett.*, vol. 28, pp. 720–721, Apr. 1992.
[62] ——, "A neural network based F0 synthesizer for Mandarin text-to-speech system," in *Proc. IEEE Visual Image Signal Processing*, Dec. 1994, vol. 141, pp. 384–390.
[63] C. H. Wu, C. H. Chen, and S. C. Juang, "An CELP-based prosodic information modification and generation of Mandarin text-to-speech," in *Proc. ROCLING VIII*, 1995, pp. 233–251.
[64] J. Choi, H. W. Hon, J. L. Lebrun, and S. P. Lee, "Yanhui, a software based high performance Mandarin text-to-speech system," in *Proc. ROCLING VII*, 1994, pp. 35–50.
[65] K. J. Chen "The identification of thematic roles in parsing Mandarin Chinese," in *Proc. ROCLING II*, 1989, pp. 121–146.
[66] L. L. Chang *et al.*, "Part of speech (POS) analysis on Chinese language," Tech. Rep., Inst. Inform. Sci., Academia Sinica, Taiwan, R.O.C., 1989.
[67] S. H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Trans. Commun.*, vol. 38, pp. 1317–1320, Sept. 1990.
[68] Y. R. Wang and S. H. Chen, "Tone recognition of continuous Mandarin speech assisted with prosodic information," *J. Acoust. Soc. Amer.*, vol. 96, pp. 2637–2645, Nov. 1994.
[69] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, 1972.

Sin-Horng Chen (S'81–M'83–SM'94) received the B.S. degree in communication engineering and the M.S. degree in electronics engineering from National Chiao Tung University (NCTU) , Taiwan, R.O.C., in 1976 and 1978, respectively, and the Ph.D. degree in electrical engineering from Texas Tech University, Lubbock, in 1983.

From 1978 to 1980, he was an Assistant Engineer for Telecommunication Laboratories, Chung-Li, Taiwan. He became an Associate Professor and a Professor at the Department of Communications Engineering, NCTU, in 1983 and 1990, respectively. He was Department Chairman from August 1985 to July 1988 and from October 1991 to July 1993. His major research area is speech processing, especially interested in Mandarin speech recognition and text-to-speech.

Shaw-Hwa Hwang received the B.S. and M.S. degrees in communication engineering and the Ph.D. degree in electronic engineering from National Chiao Tung University, Hsinchu, Taiwan, R.O.C., in 1989, 1991, and 1996, respectively.

He became an Assistant Professor at the Department of Information Management, Ming Shin Institute of Technology, Hsinchu, in 1997. His major research area is speech processing, especially Mandarin text-to-speech.

Yih-Ru Wang received the B.S. and M.S. degrees from the Department of Communication Engineering, National Chiao Tung University (NCTU), Hsinchu, Taiwan, R.O.C., in 1982 and 1987, respectively, and the Ph.D. degree from the Institute of Electronic Engineering, NCTU, in 1995.

He was an instructor of the Department of Communication Engineering, NCTU, from 1987 to 1995. In 1995, he became an Associate Professor. His general research interests are Mandarin speech recognition and the application of neural networks in speech processing.