Tone recognition of continuous Mandarin speech assisted with prosodic information

Yih-Ru Wang and Sin-Horng Chen
Department of Communication Engineering and Center for Telecommunications Research, National Chiao Tung University, Hsinchu, Taiwan 300, Republic of China

(Received 14 October 1993; revised 8 April 1994; accepted 8 July 1994)

In this paper, a simple recurrent neural network (SRNN) is employed to model the prosody of continuous Mandarin speech to assist tone recognition. For each syllable in continuous speech, several acoustic features carrying prosodic information are extracted and taken as inputs to the SRNN. If proper linguistic features extracted from the context of the syllable are set as output targets, the SRNN can learn to represent the prosodic state of the utterance at the syllable using its hidden nodes. Outputs of the hidden nodes then serve as additional recognition features to assist recognition of the tone of the syllable. The performance of the proposed tone recognition approach was examined by simulation on a multilayer perception (MLP)-based speaker-dependent tone recognition task. The recognition rate was improved from 91.38% to 93.10%. The SRNN prosodic model is further analyzed to exploit the linguistic meaning of prosodic states. By vector quantizing the outputs of the hidden nodes of the SRNN, a finite-state automata that roughly represents the mechanism of human prosody pronunciation can be obtained.

PACS numbers: 43.72.Bs

INTRODUCTION

In Mandarin speech, each written character is pronounced as a monosyllable with a tone. Monosyllables are distinguished not only by their phonemic constituents but also by their tones. This means that monosyllables with the same phonemic structure may have different meanings specified by their tones. Speech recognition for 1300 Mandarin monosyllables can therefore be conveniently decomposed into the recognition of 411 base syllables and lexical tone recognition. For the case of isolated syllable recognition, tone recognition can be run in parallel with base-syllable recognition. But in continuous Mandarin speech recognition, tone recognition usually follows base-syllable recognition so that syllable boundaries are determined before recognition features are extracted from the fundamental frequency ($F0$) and energy contours of syllables.

There are only five lexical tones in Mandarin speech: the high-level, midrising, midfalling-rising, high-falling, and neutral tones. For simplicity, these tones are commonly labeled in sequence from tone 1 to tone 5. The tone of a Mandarin monosyllable is mainly characterized by the shape of its $F0$ contour. Linguistics researchers therefore call Mandarin Chinese a "contour-tone" language (Chao, 1968). A previous study (Chao, 1968) concluded that the $F0$ contour of each of the first four tones can be represented by a single standard pattern, as shown in Fig. 1. The pronunciation of tone 5, on the other hand, is usually highly context dependent, so its $F0$ contour shape is relatively arbitrary. Tone 5 is always pronounced short and light, however.

In tone recognition of isolated Mandarin syllables, only the first four tones need to be recognized because syllables with tone 5 are rare. In the past, many recognizers have been introduced for isolated Mandarin monosyllable tone recognition based on discriminating the $F0$ contour patterns of syllables. A recognition rate of 94% has been achieved by using a multilayer perceptron (MLP)-based tone recognizer (Chang *et al.*, 1990).

But tone recognition of continuous Mandarin speech is much more complicated, because the $F0$ contour of a syllable in continuous speech is subject to various modifications. First, both its shape and its level may be seriously affected by the tones of neighboring syllables. This effect is generally known as sandhi rules (Chao, 1968; Lee *et al.*, 1989). Second, coarticulation with neighboring syllables may bring about further modifications. This is especially true when adjacent tones are of different $F0$ values. Third, the $F0$ level will be adjusted to conform to the intonation pattern of the sentence. For example, the $F0$ contour of a declarative utterance usually declines gradually. This is known as the declination effect (O'Shaughnessy and Allen, 1983; Lee *et al.*, 1989). Last, the $F0$ level will also be seriously affected by the prosody of the utterance. This is the major effect that is studied in this paper.

In the past few years, several researchers have investigated tone recognition of continuous Mandarin speech. Wang (Wang, 1988) used a DHMM-based approach to recognize tones for disyllabic and trisyllabic words. Wang *et al.* (Wang *et al.*, 1990) proposed a rule-based approach that considers both the declination effect and the accent effect based on Fujisaki's model (Fujisaki *et al.*, 1988). A recognition rate of 93% was achieved for tone recognition of four-syllable idioms. Wang and Chen (Wang and Chen, 1994) proposed a tone recognizer that uses context-dependent HMM models to compensate for the effect of sandhi rules and used a two-level network structure to model the declination effect of sentential utterances. Wang and Chen (Wang and Chen, 1993) introduced an MLP-based approach to compensate for both the coarticulation effect and sandhi rules caused by neighboring syllables by directly incorporating contextual

 0001-4966/94/96(5)/2637/9/$6.00

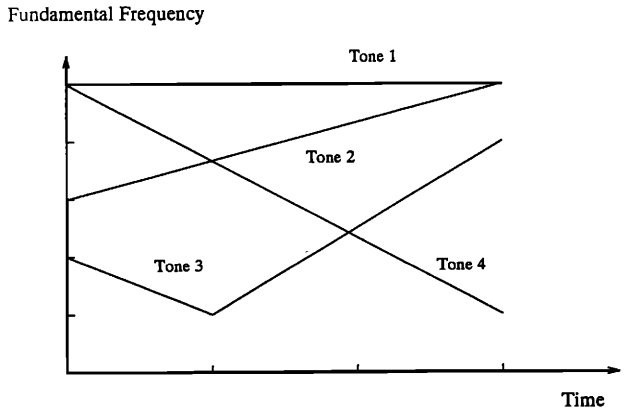
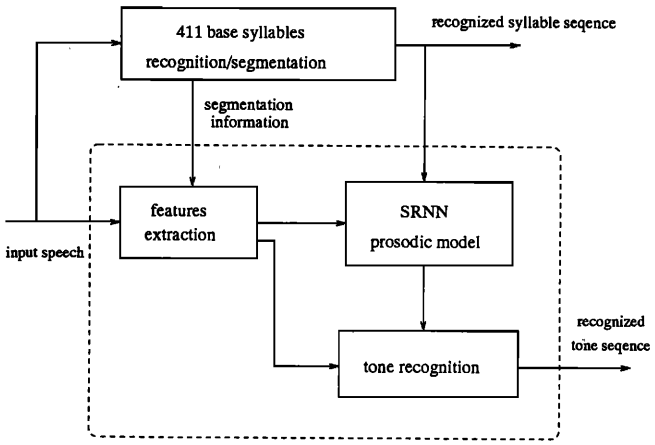


FIG. 1. Standard $F0$ contour patterns of the first four tones.



The proposed tone recognizer

FIG. 2. The hierarchical structure of the $F0$ contour of a sentential utterance.

features as additional input features to aid in tone recognition. Although promising results were obtained in these studies, high-level factors such as prosodic information and syntactic and semantic features were still not properly considered. Since the $F0$ contours of syllables are seriously affected by such high-level factors, we believe that modeling these factors will surely help tone recognition. This motivates our attempt to use a prosodic model to assist tone recognition.

Continuous speech includes suprasegmental information such as stress, intonation pattern, and timing structure (tempo). This information is generally referred to as the prosody of the speech, which in turn is affected by the sentence type, the syntax structure, semantics, and the emotion and encompassing attitude of the speaker. According to a previous study (Lea, 1980), the prosody of a speech is a dominating factor that determines the energy level, the length of silence between syllables, the duration of the vowel, and the $F0$ level of syllables. As far as tone recognition is concerned, the prosodic effects on the $F0$ and the energy contours of a syllable will be superimposed on the tonal $F0$ from the syllable and the neighboring syllables (O'Shaughnessy and Allen, 1983). Figure 2 illustrates the hierarchical structure of these effects on the $F0$ contour for a declarative utterance. First, the global intonation pattern at the utterance level is shown in Fig. 2(a). Then, at the phrase or clause level, a local declination is shown in Fig. 2(b). Finally, within a phrasal segment, each syllable has its own $F0$ pattern, as shown in Fig. 2(c). This $F0$ modulation phenomenon in Mandarin speech is like a ripples-on-the-wave pattern (Chao, 1968). Obviously, high-level effects will destort the $F0$ pattern of a tone, causing it to deviate from its standard pattern and therefore hampering tonal discrimination. A good prosodic model is expected to compensate for the prosodic effect and improve tone recognition.

In this paper, a neural network-based approach is adopted to model the prosody of continuous Mandarin speech so as to improve tone recognition. The basic idea is to employ a simple recurrent neural network (SRNN) to infer prosodic states of syllables from acoustic features. The SRNN prosodic model can be regarded as a system that identifies articulary mechanisms of prosody from speech. Figure 3 shows a schematic diagram of the SRNN prosodic model. The SRNN is used to learn the relationship between the sequence of acoustic features carrying prosodic information extracted from the input utterance and the linguistic features extracted from the text associated with the utterance. Through training with a large set of utterances with texts, rules for inferring prosodic states from acoustic features can be automatically learned and implicitly memorized by the SRNN. Hidden nodes of the SRNN are organized to represent the prosodic state of the utterance at the syllable being processed. We can therefore take outputs of these hidden nodes as additional recognition features to assist in tone discrimination

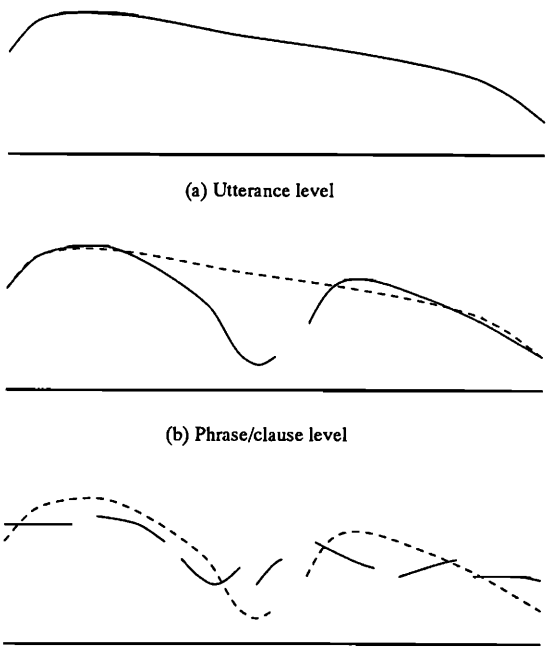


FIG. 3. The block diagram of the Mandarin speech recognition system assisted with the proposed prosodic model.

This paper is organized as follows. Section I describes the speaker-dependent speech database used in this study. Section II presents the proposed SRNN prosodic model. Section III evaluates the effectiveness of the proposed prosodic model. An MLP-based tone recognizer is adopted in this study. Finally, conclusions are stated in the Sec. IV.

I. SPEECH DATABASE

A continuous Mandarin speech database read by a single male speaker was used in this study. The database consists of two parts. The first part contains 240 short sentences. The second part contains 103 newspaper paragraphs. The number of syllables in an utterance in the first part ranges from 5–81, including punctuation marks, and the number in the second part ranges from 30–431. All utterances were spoken naturally at a speed of 3.5–4.5 syllables per s. Note that the speaking rate influences the $F0$ patterns of tones. The $F0$ contour shapes follow their standard tone patterns in slow speech, but they may become seriously distorted when the speaking rate exceeds 4.5 syllables per s. The database contains, in total, 18960 syllables and 1971 punctuation marks. The database is composed of 1196 phonetically different syllables out of the 1319 possible syllables in Mandarin.

All speech signals were digitally recorded with a 20-kHz sampling rate. They were then divided into 4-ms frames and manually segmented into silence, unvoiced, and voiced parts based on waveform, energy, zero crossing rate, LPC coefficients, cepstrum, and delta cepstrum. Three acoustic features, fundamental frequency ($F0$), log energy, and zero-crossing rate, were then extracted from the down-sampled 10-kHz speech signal for prosodic and tone information analysis. The $F0$ contour was estimated by using the simple inverse filter tracking (SIFT) algorithm (Markel and Gray, 1972) with manual error correction. The window size for $F0$ analysis was 40 ms with 10-ms window shift, and the window size for both log-energy and zero-crossing rate analysis was 20 ms.

Finally, all texts were segmented into lexical words using a Chinese lexicon (supplied by a local computer company working on machine translation). Words in the lexicon are one to five syllables long. The text was then further processed manually to find proper nouns and words missing in the lexicon. After all word sequences were obtained, linguistic features were extracted from each syllable to be used as targets for training the SRNN.

II. THE SRNN-BASED PROSODIC MODEL

An SRNN was used in this study to model the prosody of Mandarin speech. The motivation of using an SRNN-based approach follows the work of Elman (Elman, 1990), in which an SRNN was employed to model a hidden mechanism generating an infinite corpus of data sequences. The structure of the SRNN used in this study is shown in Fig. 4. It is a three-layer network with one hidden layer. All outputs of a hidden layer are fed back to the input layer with unit-time delays. The output of each neuron is a nonlinear function of the weighted sum of the input signals. Let the output of the kth neuron in the ith layer be denoted by $O_k^{(i)}(n)$ and

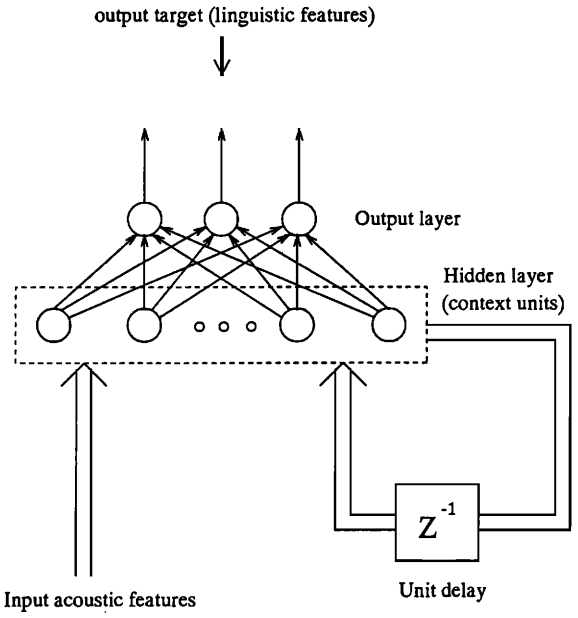


FIG. 4. The architecture of the SRNN prosodic model.

the connection weight from the jth neuron in the i_1^{th} layer to the kth neuron in the i_2^{th} layer be denoted by $W_{k,j}^{(i_2,i_1)}$, where n is the time index. Then,

$$O_k^{(3)}=f(\text{net}_k^{(3)}(n)),$$

$$\text{net}_k^{(3)}(n)=\sum_j W_{k,j}^{(3,2)}O_j^{(2)}(n),$$

$$O_j^{(2)}=f(\text{net}_j^{(2)}(n)),$$

$$\text{net}_j^{(2)}(n)=\sum_l W_{j,l}^{(2,1)}I_l(n)+\sum_{l'} W_{j,l'}^{(2,2)}O_{l'}^{(2)}(n-1), \quad (1)$$

where $I_l(n)$ represents the input signals and f is a sigmoid function defined as

$$f(x)=\frac{1}{1+e^{-x}}. \quad (2)$$

The SRNN has the ability to model finite state machines and to simultaneously represent similarity and difference between several sequences (Servan-Schreiber *et al.*, 1991; Elman, 1991). As pointed out by Rumelbert (Rumelbert *et al.*, 1987), the pattern of activations on the hidden units corresponds to an "encoding" or "internal representation" of the input sequence. The hidden units are called the context units. In this study, the SRNN is employed to explore the internal representation of a sequence of acoustic feature vectors which convey prosodic information. The embedded prosodic states of the utterance are expected to be sequentially captured by the SRNN and implicitly represented by the context units. Outputs of the context units are then used to assist tone recognition. In the following, the SRNN prosodic model is discussed in more detail.

First, let us discuss the selection of proper input acoustic features. It is known from previous studies that many acoustic features are affected by the prosody of speech (Shimo-

TABLE I. The recognition results of the SRNN prosodic model (unit: %).

	Recognition result		
	Intraword	Interword	P.M.
Intraword	73.95	25.90	0.14
Interword	23.35	72.89	3.79
P.M.	1.12	7.23	91.65

daira and Kimura, 1992; Hieronymus *et al.*, 1992). For instance, the level and the dynamic range of the $F0$ contour and the energy level of a vowel are related to the stress level of a syllable; the energy dip and the duration of silence between syllables are related to the tempo and rhythm; and the duration of a vowel is related to both tempo and stress. All these acoustic features can be used as input features of the SRNN to model the prosody. In the bottom-up tone recognition task of this study, only some basic acoustic features carrying prosodic information were used. These were: (1) the log-energy mean and (2) the duration of the silence between the processing syllable and the following syllable; and (3) the normalized log-energy mean, (4) the lengthening factor (Price *et al.*, 1991), and (5) the $F0$ mean of the vowel of the processing syllable. Among these features, the silence duration is an important cue to indicate the degree of conjunction between two consecutive syllables and is normalized by the type of initial in the following syllable. The normalization is used to compensate for the large variation in silence duration due to different types of initial of the following syllable. For instance, a longer silence duration always exists before a plosive initial. The lengthening factor is defined as the normalized duration of the vowel. It carries some syntactic information, such as clause and phrase boundaries (Price *et al.*, 1991). The log-energy mean and the $F0$ mean of the vowel carry information about the intonation of the utterance and the stress pattern of the phrase.

The three low-level linguistic features are selected as the target function of the SRNN, which specify whether an intraword, inter-word, or punctuation mark (PM) follows the processing syllable. According to the criterion of minimizing the mean-square error between actual outputs and desired output targets, the SRNN can be trained by the recursive back propagation learning rule for recurrent neural networks (Lee *et al.*, 1991).

All the data in the database described above were used to train the SRNN. There are a total of 18617 output target vectors in the database, including 7675, 9158, and 1784 vectors for intra-word, inter-word, and punctuation mark indicators, respectively. The number of context units was empirically set to be 25. Table I shows the intra-word/inter-word/PM recognition results. It can be seen from Table I that the recognition rate for punctuation marks is very high. Most errors in recognizing punctuation marks resulted from marks that do not cause long silences in the pronunciation, such as quotation marks and colons. The recognition rates for both interword and intraword detection were around 73%. This shows that many tokens of interword and intraword are indistinguishable. Nevertheless, using word segmentation informations causes no harm to our mission because the outputs of the hidden layer rather than the outputs of the output layer are used as additional recognition features in the following tone recognition tests.

A. Analysis of the prosodic model

In order to identify the characteristics of the inferred prosodic model, we shall now analyze the SRNN in detail by examining the prosodic state sequence stored in the context units of the SRNN. By clustering all output vectors of the context units into a finite number of sets and associating each set with a state (Servan-Schreiber *et al.*, 1991), we can obtain a finite state automata to roughly display the mechanism of prosody. Figure 5 depicts the resulting 8-state prosodic model in which only the first two or three most significant state transitions are drawn. Detailed initial state probabilities and state transition probabilities are listed in Table II. If the prosodic state sequences produced by the model for some training utterances are examined closely, a linguistic interpretation of these eight quantized states can be found. States 3 and 8 always occur at the beginning of a sentential utterance or a clause. State 3 may also appear at the beginning of a major phrase. State 6 is associated with the ending syllable of a sentential utterance. State 2 is usually associated with the ending syllable of a polysyllabic phrase preceding a long silence. State 4 is usually associated with the beginning of a minor phrase or word that follows a short silence. States 1 and 5 usually follow state 3 or 8. State 5 is also often associated with the ending syllable of a word. State 7 is an unimportant state which appears relatively infrequently. Moreover, states 1 and 7 are associated with intermediate syllables of phrases or words. Figure 6 shows a typical example of the state sequence produced by the SRNN prosodic model for a paragraph. The utterance consists of 41 syllables and two punctuation marks: "ying-1 jiun-1" (British army) "fa-1 yian-2 ren-2" (spokesman) "ou-1 uen-2" (name of the spokesman) "tze-2" (on the other hand) "jeng-4 shy-2" (prove) "," (comma) "i-2 jia-4" (one classifier) "ying-1 jiun-1" (British army) "jy-2 sheng-1 ji-1" (helicopter) "tsuei-2 huei-3" (destroy) "i-1 la-1 ke-4" (Iraq) "hai-3 jiun-1" (navy) "i-4 shou-1" (one classifier) "pei-4 bei-4" (fitted with) "u-3 ting-3" (five classifier) "ji-1 pau-4" (machine gun) "de-5" (particle) "su-1 lian-2 jy-4" (made in Russia) "kuai-4 su-4" (fast) "shiun-2 luo-2 ting-3" (patrol boat) "." (period). The utterance starts with state 8. The first sentence ends at the tenth syllable, associated with state 6. The second sentence starts at the eleventh syllable, associated with state 3. Other syllables with state 3 are usually located at the beginning of major phrases. Most ending syllables of phrases are associated with state 2. Some minor phrases start at syllables associated with state 4. We can conclude that many prosody characteristics have indeed been inferred by the SRNN.

III. TONE RECOGNITION ASSISTED WITH THE PROSODIC MODEL

We now examine the effectiveness of the prosodic model in improving the performance of tone recognition by simulating a speaker-dependent continuous-speech tone recognition task. The database described previously was first

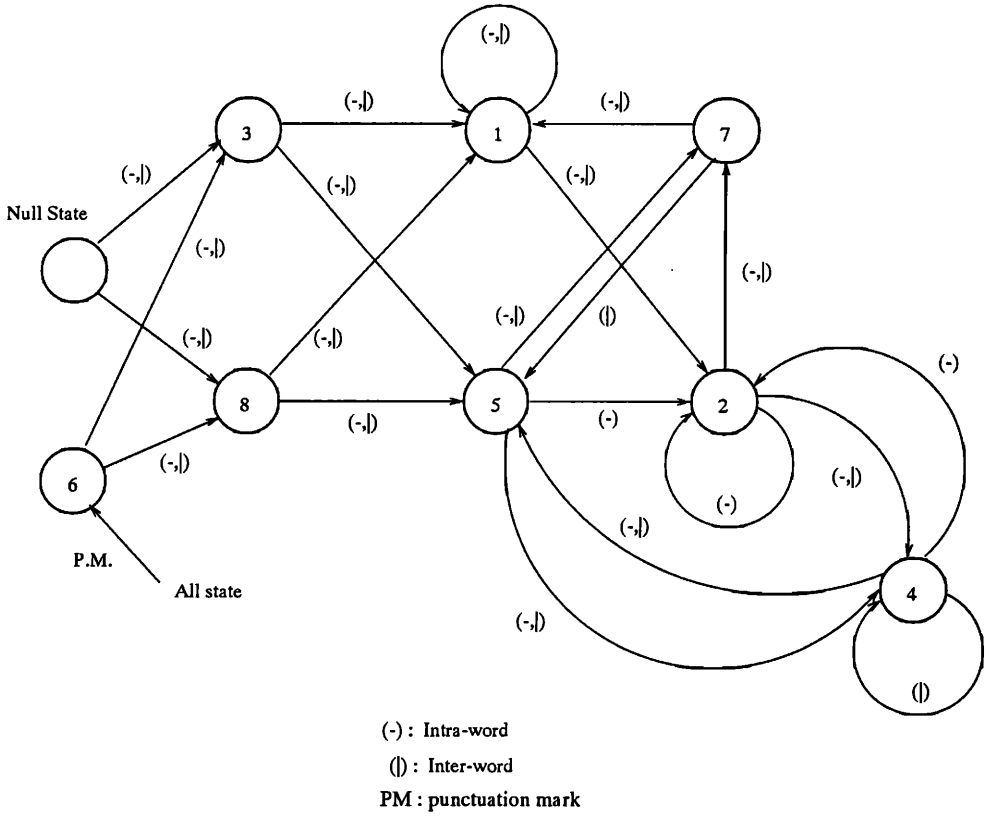


FIG. 5. The finite state automata built by vector quantizing the context units of the SRNN prosodic model.

preprocessed to relabel some syllables of tone 3 as tone 2 that merges sequences of tone 3-tone 3 with tone 2-tone 3 by carefully listening examination. Some tone 3's are actually tone 2's as the result of a tone sandhi ruler (Chao, 1968). The database was then split into two sets, one for training and the other for testing. The training set consisted of 287 utterances with 15 953 syllables and the test set consisted of 56 utterances with 3015 syllables. The distributions of the five tones in these two sets is listed in Table III.

A. The basic MLP-based tone recognizer

The basic tone recognizer used in this study was an LR context-dependent, neural net-based tone recognizer (Wang and Chen, 1993) which uses some local acoustic features as input recognition features. The network structure was a feed-forward multilayer perceptron (MLP) with a single hidden layer. There are a total of 36 input features used in the basic tone recognizer, including 10 features extracted from the processing syllable, 16 contextual features, and 10 binary features representing the tones of the two nearest neighboring syllables. The 10 features extracted from the processing syllable are the duration of the $F0$ contour, means of three uniformly divided log-energy subcontours, and means and slopes of three uniformly divided $F0$ subcontours (Chang *et al.*, 1990). The 16 contextual features include (1) three features (i.e., log-energy, $F0$ mean and slope) extracted from the last segment of the preceding syllable, (2) three features extracted from the first segment of the following syllable, (3) duration of the silence as well as (4) log energies, zero crossing rates, and durations of the unvoiced segments located before and after the processing syllable, and (5) two binary indicators showing whether the processing syllable is the first or the last syllable of the testing utterance. Among these 16 contextual features, the first 6 features in (1) and (2) are the primary features for coping with the effect of neighboring $F0$ contours. The following two features in (3) are used to implicitly represent the tightness of relations between the processing syllable and the two nearest neighbors. The tones of the two nearest neighboring syllables are also used because the $F0$ contour shape of the processing syllable may be seriously affected by them due to the sandhi rule and the coarticulation effect. Because the tones of neighboring syllables are either not known in advance or can only be estimated from previous recognition, tone recognition tests for syllables in an input utterance cannot be done independently. A recognition procedure based on the decision rule of minimal total risk is employed here to simultaneously recognize tones of all syllables in a sentential utterance. The steps of the recognition procedure are as follows. First, rather than directly taking the MLP as a tone recognizer, we regard it as a mechanism for calculating the risk of each tone-trigram composed of the tones of the processing and the two nearest neighboring syllables. Second, we define an objective function for each candidate tone sequence for the whole input utterance by accumulating the risks of all tone-trigrams in the tone sequence. Specifically, given the feature vector sequence, $(\mathbf{X}(j))_{j=1,N}$, of the input utterance with N syllables, the objective function for the candidate tone sequence $(T(j))_{j=1,N}$ is defined as

TABLE II. The initial and transition probabilities of the 8-state automata built from the SRNN prosodic model. (a) Initial probability: Pr(state of the first syllable of an utterance). (b) Transition probability: Pr(from state i to state j).

(a)	State	1	2	3	4	5	6	7	8
	Intra	0	0.03	0.70	0	0	0	0	0.27
	Inter	0	0.04	0.64	0.01	0	0	0	0.34
	P.M.	0	0	0	0	0	1.00[a]	0	0

(b)	State i		State j 1	2	3	4	5	6	7	8
		Intra	0.28	0.43	0.06	0.05	0.12	0.03	0.08	0.03
	1	Inter	0.47	0.23	0.18	0.15	0.05	0	0.08	0.01
		P.M.	0.02	0.10	0	0	0.07	0.87	0.07	0
		Intra	0.08	0.25	0.23	0.36	0.02	0.07	0.02	0.04
	2	Inter	0.02	0.07	0.30	0.54	0	0	0.01	0.06
		P.M.	0	0.04	0.01	0.01	0	0.94	0	0
		Intra	0.42	0.14	0	0.01	0.35	0.01	0.07	0
	3	Inter	0.59	0.07	0.01	0.02	0.30	0	0.09	0
		P.M.	0.06	0.18	0	0	0.03	0.73	0	0
		Intra	0.04	0.27	0.01	0.17	0.35	0.09	0.07	0.03
	4	Inter	0.08	0.11	0.04	0.44	0.28	0	0.04	0.01
		P.M.	0	0.02	0	0	0	0.97	0	0
		Intra	0.06	0.30	0.06	0.37	0.06	0.08	0.03	0.03
	5	Inter	0.06	0.09	0.12	0.66	0.01	0	0.02	0.04
		P.M.	0	0.04	0	0	0.07	0.96	0	0
		Intra	0.02	0.10	0.66	0.06	0	0.01	0.01	0.14
	6	Inter	0.02	0.02	0.70	0.07	0	0	0	0.19
		P.M.	0	0.19	0.22	0	0	0.52	0	0.07
		Intra	0.10	0.19	0.01	0.02	0.64	0.02	0.03	0
	7	Inter	0.33	0.13	0.03	0.03	0.37	0.01	0.08	0.02
		P.M.	0	0.12	0	0	0.06	0.82	0	0
		Intra	0.35	0.09	0	0	0.49	0	0.06	0
	8	Inter	0.49	0	0	0.01	0.39	0.01	0.11	0
		P.M.	0.17	0.06	0	0	0.11	0.67	0	0

[a]Only 2 P.M. were found.

$$R((T(j))_{j=1,N}|(\mathbf{X}(j))_{j=1,N}) = \sum_{j=1}^{N}\sum_{i=1}^{5}\{(O_i(\mathbf{X}(j),T(j-1),T(j+1))-t_i(j))^2\}. \quad (3)$$

Here, the output of the neural network, $O_i(\mathbf{X}(j),T(j-1),T(j+1))$, is a function of the tone trigram $(T(j-1)$, tone i, $T(j+1))$; and the desired output, $t_i(j)$, of the network for the jth syllable is given by

$$t_i(j)=\begin{cases}1, & \text{if } T(j)=\text{tone } i\\ 0, & \text{if } T(j)\neq\text{tone } i.\end{cases} \quad (4)$$

The tone recognition test then becomes the problem of finding the best tone sequence $(\hat{T}(j))_{j=1,N}$ that minimizes the objective function. In implementation, this function can be efficiently solved by dynamic programming. We note that an extra linguistic constraint is added in the search for the best tone sequence for an input utterance to inhibit the tone of the first syllable from being tone 5 because this can never occur in natural Mandarin speech. The performance of the basic MLP tone recognizer was first examined by simulations using the database described earlier. A benchmark recognition rate of 96.07% and 91.11% for inside and outside tests was achieved when 80 hidden neurons was used.

B. Tone recognition assisted with the SRNN prosodic model

We now incorporate the SRNN prosodic model into the basic MLP recognizer to assist tone recognition. The hidden units of the SRNN prosodic model were fed into the MLP tone recognizer as additional input features via a concentration layer. The concentration layer is used as a buffer to relieve the expansion on the input dimension of the MLP tone recognizer. The number of neurons in the concentration layer was empirically chosen to be 4. The block diagram of the new tone recognizer is shown in Fig. 7. We note that special treatment must be given to the last syllable of each utterance because there are no prosodic states associated with it. Since fact that state 6 roughly represents the ending of a sentential utterance, we take the average output vector of context units of all instances of state 6 in the training set as the prosodic state of the last syllable. A recognition rate of 97.07% and 92.80% for inside and outside tests was achieved when 110 hidden neurons were used. Compared with the basic tone recognizer, 19% of the errors in the outside test were corrected.

In order to check the relation between prosodic states and tone recognition errors, the error rates in the eight quantized prosodic states were calculated (see Fig. 8). All of the errors in both the inside and the outside tests were included.

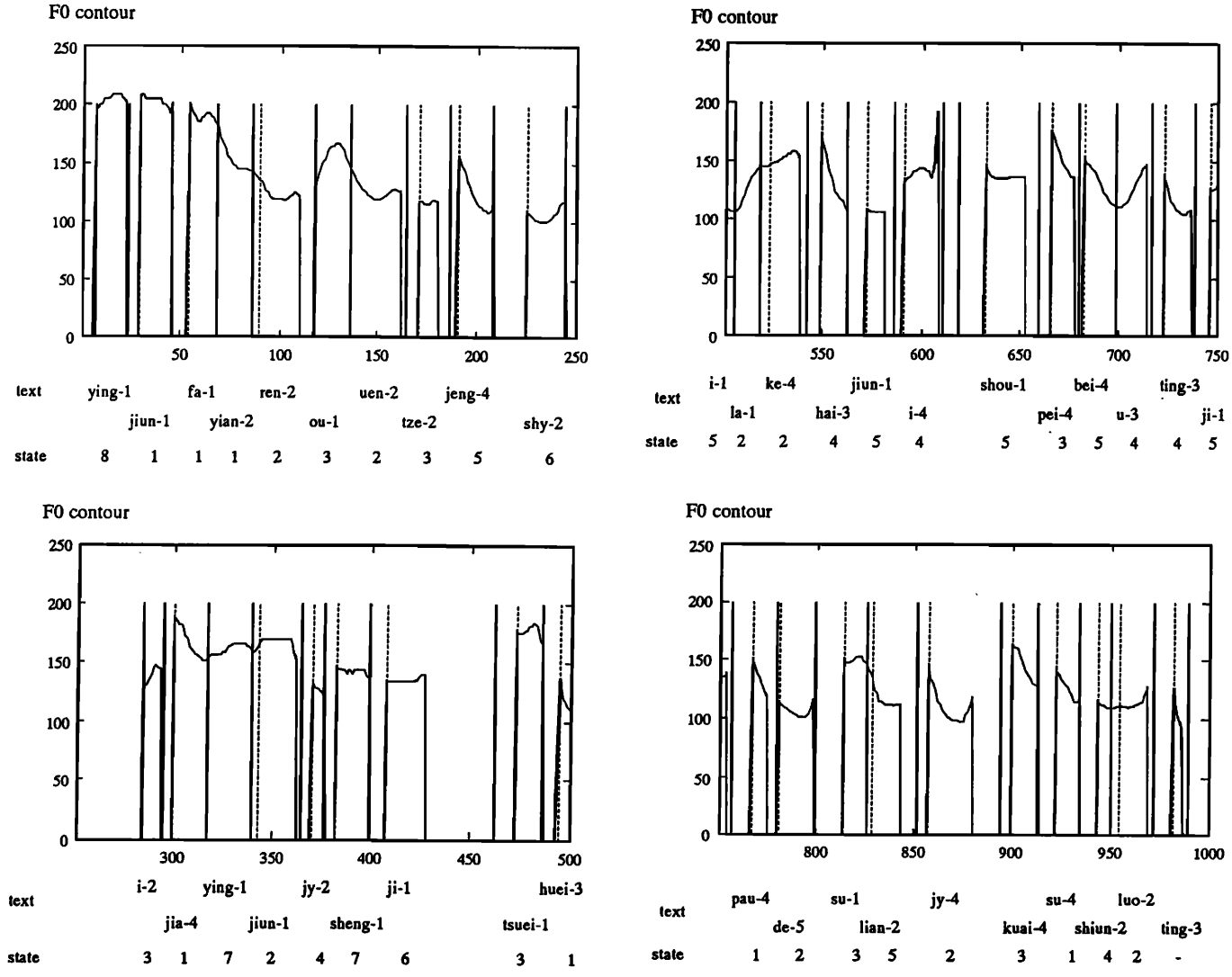


FIG. 6. An example of a prosodic state sequence produced by the SRNN prosodic model for an input paragraph.

It can be seen from Fig. 8 that the error rates in all prosodic states were improved when the prosodic model was used to assist tone recognition. The improvement was more significant for state 3, state 4, and state 8. As discussed above, these three states correspond to the beginning of sentences and phrases. Variations in the $F0$ level in these three states are larger and hence potentially hamper the tone recognition. This result shows that the SRNN prosodic model can capture the prosody of speech so as to partially compensate for the effect of large $F0$ variation on tone recognition.

Finally, the tone recognizer was fine tuned by using a generalized probabilistic descent (GPD) algorithm (Katagirl *et al.*, 1991) to discriminatively adjust all its parameters. In the GPD algorithm, the outputs of the MLP tone recognizer are taken as discriminant functions of the five tone classes. A misclassification measure for the jth syllable is then defined as

$$d(j) = -\text{net}_i^{(3)}(j) + \max_{n \neq i} \text{net}_n^{(3)}(j), \tag{5}$$

to judge the goodness of the current decision. Here tone i is the correct tone class of the jth syllable. Then a loss function for evaluating the cost of the current decision for the jth syllable is defined as

$$l(j) = f(d(j)), \tag{6}$$

where f is a sigmoid function. The objective can then be formulated as an optimization procedure to find the best tone sequence, $(\hat{T}(j))_{j=1,N}$, that minimizes the summation of the loss functions of all syllables in each input training utterance

TABLE III. The distributions of the five tones in the training and the test data sets.

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Total
Training set	3358	3655	2935	5277	720	15 953
Test set	707	645	516	1036	109	3015

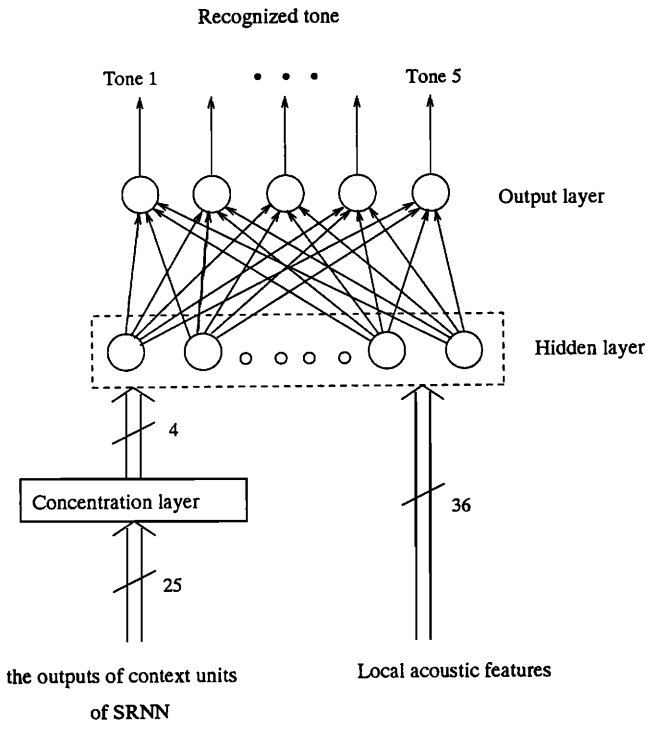


FIG. 7. The block diagram of the MLP tone recognizer assisted with the SRNN prosodic model.

$$R((T(j))_{j=1,N}|(\mathbf{X}(j))_{j=1,N}) = \sum_{j=1}^{N} l(j). \quad (7)$$

The GPD algorithm is an iterative procedure that adjusts all parameters of the MLP recognizer using the steepest descent method to realize the optimization procedure. Initialized with the parameters obtained in the previous tone recognition test, the MLP tone recognizer can be fine tuned by the GPD algorithm with the goal of minimizing an approximation of the tone recognition errors. Recognition rates of 97.73% and 93.10% were achieved for the inside and the outside tests,

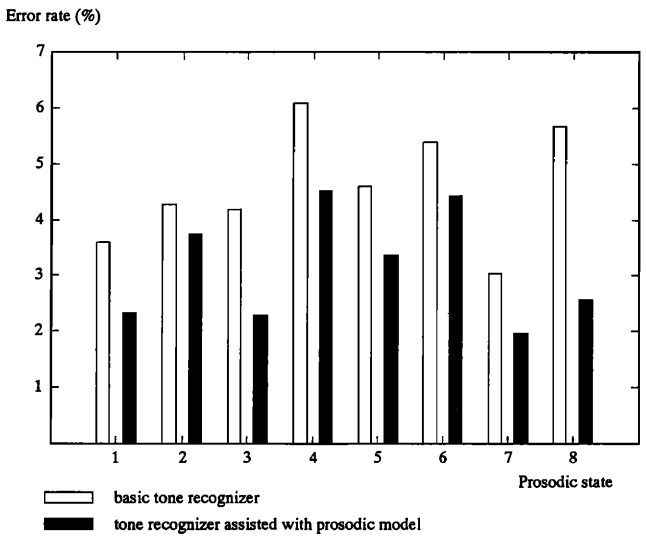


FIG. 8. Error rates of tone recognition for different prosodic states.

TABLE IV. Confusion table for tone recognition using the recognizer assisted with the prosodic model (unit: %).

	Recognition result				
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5
Tone 1	93.2	4.5	0.0	1.9	0.4
Tone 2	3.2	91.9	3.1	0.6	1.3
Tone 3	0.4	4.6	90.7	1.5	2.6
Tone 4	1.2	0.7	1.4	96.2	0.6
Tone 5	0.9	7.3	9.1	1.8	80.9

respectively. The confusion table for the outside test is given in Table IV. It can be seen from the table that the recognition rate of tone 5 is still far below average. Though detailed error analysis, we found that the most probable error for tone 5 is to recognize it as tone 3. This error mainly resulted from the similarity of the $F0$ contour patterns of many tone 5 syllables to the standard tone pattern of tone 3. In fact, the $F0$ contour pattern of tone 5 is generated by using a shorter version of the tone 3 pattern in Lee's Mandarin text-to-speech system (Lee *et al.*, 1989). Finally, for a performance comparison, the basic MLP tone recognizer was also discriminatively trained by the GPD algorithm. Recognition rates of 96.38% and 91.38% were obtained for the inside and the outside tests, respectively. So, for the outside test, the proposed prosodic model reduces errors by about 20% overall.

IV. SUMMARY

In this paper, an SRNN-based approach to modeling the prosody of Mandarin speech from acoustic features has been studied. By using acoustic features carrying prosodic information as inputs and setting some simple linguistic features as output targets, we can train the SRNN to learn the prosodic characteristics of Mandarin speech. Experimental results confirm that the prosody states of an input utterance can be sequentially captured by the SRNN model and implicitly represented by the hidden units. By incorporating the SRNN prosodic model into an MLP tone recognizer to assist tone recognition, we improved a Mandarin tone recognition rate from 91.38% to 93.10%.

ACKNOWLEDGMENTS

This work was supported by the National Science Council (under Contract No. NSC83-0404-E009-091) and Telecommunication Laboratories, MOTC, Taiwan, Republic of China. The speech database was supplied by Telecommunication Laboratories.

Chang, P. C., Sun, S. W., and Chen, S. H. (**1990**). "Mandarin tone recognition by mult-layer perceptron," Proc. 1990 IEEE Conf. Acoust. Speech Signal Process. Vol. 1, 517–520.

Chao, Y. R. (**1968**). *A Grammar of Spoken Chinese* (Univ. of California, Berkeley California).

Elman, J. L. (**1990**). "Representation and Structure in Connectionist Models," in *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (MIT, Cambridge, MA).

Elman, J. L. (**1991**). "Distributed representations, simple recurrent networks, and grammatical structure," Mach. Learn. **7**, 195–225.

Fujisaki, H., *et al.* (**1988**). "Application of *F*0 contour command-response model to Chinese tones," Reports of Autumn meeting, J. Acoust. Soc. Jpn. 197–198.

Hieronymus, J. L., McKelvie, D., and McInnes, F. R. (**1992**). "Use of acoustic sentence level and lexical stress in HSMM speech recognition," Proc. 1992 IEEE Conf. Acoust. Speech, Sig. Process. Vol. 1, 225–227.

Katagirl, S., Lee, C. H., and Hwang, B. H. (**1991**). "Discriminative multilayer feed-forward networks," Proc. of 1991 IEEE workshop on Neural Networks Sig. Process. 11–20.

Lea, W. A. (**1980**). *Trends in Speech Recognition, chapter 8: Prosodic Aids to Speech Recognition* (Prentice-Hall, New York, 1980).

Lee, L. S., Tseng, C. Y., and Ouh-Young, M. (**1989**). "The synthesis rules in a chinese text-to-speech system," IEEE Trans. Acoust. Speech Sig. Process. **37**, 1309–1320.

Lee, S. J., Kim, K. C., Yoon, H., and Cho, J. W. (**1991**). "Application of fully recurrent neural network for speech recognition," Proc. 1991 IEEE Conf. Acoust. Speech Sig. Process. Vol. 1, 77–80.

Markel, J. D., Gray, Jr., and A. H. (**1972**). "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio Electroacoust. **AU-20**, 367–377.

O'Shaughnessy, D., and Allen, J. (**1983**). "Linguistic modality effects on fundamental frequency in speech," J. Acoust. Soc. Am. **74**, 1155–1170.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, and Fong, C. (**1991**). "The use of prosody in syntactic disambiguation," J. Acoust. Soc. Am. **90**, 2956–2970.

Rumelbert, D. E., Hinton, G., and Williams, R. J. (**1987**). "Learning internal representations by error propagation," *Parallel Distributed Processing, Vol. 1: Foundations* (MIT, Cambridge).

Servan-Schreiber, D., Cleeremans A., and McClelland, J. L. (**1991**). "Graded state machines: The representation of temporal contingencies in simple recurrent networks," Mach. Learn. **7**, 161–193.

Shimodaira, H., and Kimura, M. (**1992**). "Accent phrase segmentation using pitch pattern clustering," Proc. 1992 IEEE Conf. Acoust. Speech Signal Process. Vol. 1, 217–220.

Wang, C. F., Fujisaki, H., and Hirose, K. (**1990**). "Chinese Four Tone Recognition based on the Model for Process of Generating *F*0 Contours of Sentences," Proc. 1990 Int. Conf. Spoken Lang. Process. Jpn., Vol. 1, 221–224.

Wang, R. D. (**1988**). "Tone recognition of continuous Mandarin speech," masters thesis, National Tsing Hua Univ., May 1988.

Wang, Y. R., and Chen, S. H. (**1993**). "Tone recognition of continuous Mandarin speech based on neural networks," *1993 Int. Symposium on Artificial Neural Networks* (National Chiao Tung Univ.), F-01–F-10.

Wang, Y. R., and Chen, S. H. (**1994**). "Tone recognition of continuous Mandarin speech based on hidden Markov model," Int. J. Pattern Recog. Artific. Intell. **8**, 233–246.