A Modular RNN-Based Method for Continuous Mandarin Speech Recognition

Yuan-Fu Liao and Sin-Horng Chen, *Senior Member, IEEE*

***Abstract*—A new modular recurrent neural network (MRNN)-based method for continuous Mandarin speech recognition (CMSR) is proposed. The MRNN recognizer is composed of four main modules. The first is a sub-MRNN module whose function is to generate discriminant functions for all 412 base-syllables. It accomplishes the task by using four recurrent neural network (RNN) submodules. The second is an RNN module which is designed to detect syllable boundaries for providing timing cues in order to help solve the time-alignment problem. The third is also an RNN module whose function is to generate discriminant functions for 143 intersyllable diphone-like units to compensate the intersyllable coarticulation effect. The fourth is a dynamic programming (DP)-based recognition search module. Its function is to integrate the other three modules and solve the time-alignment problem for generating the recognized base-syllable sequence. A new multilevel pruning scheme designed to speed up the recognition process is also proposed. The whole MRNN can be trained by a sophisticated three-stage minimum classification error/generalized probabilistic descent (MCE/GPD) algorithm. Experimental results showed that the proposed method performed better than the maximum likelihood (ML)-trained hidden Markov model (HMM) method and is comparable to the MCE/GPD-trained HMM method. The multilevel pruning scheme was also found to be very efficient.**

***Index Terms*—Mandarin speech recognition, MCE/GPD algorithms , modular recurrent neural networks.**

I. Introduction

CURRENTLY, the dominant technology for continuous speech recognition (CSR) is based on HMMs [1], [2]. HMMs are good at statistically based acoustic modeling and provide a fundamental structure flexible enough for the recognition of nonstationary speech signals. Aside from the HMM approach, the artificial neural network (ANN)-based approach is also attractive because ANNs have the distinction of possessing high discrimination ability obtained through competitive learning and hence are potentially good for speech pattern discrimination [3], [4]. Although many ANN-based methods have been proposed previously, only a few of them are suitable for CSR because of the lack of fundamental structures to deal with the time-alignment problem. Most ANN structures, such as multilayer perceptron (MLP) and RNN, are only good at discriminating short input speech patterns of several frames in length. They are therefore good pattern classifiers instead of good recognizers for nonstationary continuous speech signals.

Manuscript received June 18, 1998; revised February 25, 2000. This work was supported by the National Science Council, R.O.C., under Contract NSC87-2213-E009-027. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mazin Rahim.

The authors are with the Department of Communication Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C. (e-mail: schen@cc.nctu.edu.tw).

Publisher Item Identifier S 1063-6676(01)01326-8.

Recently, a hybrid HMM/ANN approach to CSR has attracted the attention of many researchers in the area of ANN-based speech recognition [3]–[6]. This approach aims at integrating the advantages of both ANN and HMM technologies. One popular approach [3], [6] is to employ an MLP or RNN pattern classifier to replace phone-like HMM models for computing the observation probabilities of all phones. Modeling of the temporal structure of speech signals for solving the time-alignment problem in the recognition search is still performed implicitly under the HMM framework. This approach is more efficient on modeling acoustic units and is easier to take care acoustic context by directly modeling high-dimensional input speech patterns of several frames. However, its performance has been shown, by experimental results, to be only slightly better than the ML-trained HMM method [4].

In this paper, a new hybrid DP/ANN method for continuous Mandarin speech recognition is proposed. It applies the "*divide and conquer*" principle [7], [8] of modular neural network technology [9]–[12] using prior phonetic knowledge to design a sophisticated MRNN recognizer which discriminates acoustic units instead of modeling them. The basic idea is to first divide the task of CMSR into several subtasks of discriminating smaller speech segments, then tackle them separately using expert RNN modules, and finally integrate partial solutions to solve the complete problem.

Specifically, the task of CMSR is first divided into three subtasks:

1) discrimination of 412 base-syllables (including a null one for silence);
2) detection of syllable boundaries;
3) discrimination of inter-syllable speech segments.

A sub-MRNN module and two RNN modules are then designed to tackle these three subtasks separately. The sub-MRNN is an extended version of the MRNN proposed previously for isolated Mandarin base-syllable recognition [13] and is composed of four RNN submodules with architecture conforming to the phonetic structure of Mandarin base-syllables. Its function is to generate discriminant functions for all 412 base-syllables. One of the two RNN modules is designed to detect syllable boundaries to provide explicit timing information to help the recognition search. It is referred to as the boundary-detection RNN module. Another, referred to as the inter-syllable-segment discrimination RNN module, is used to generate the discriminant functions of 143 classes of intersyllable diphone-like units to compensate the intersyllable coarticulation effect [14]–[19].

Solutions of these three subtasks are lastly combined to solve the complete task of CMSR using a DP-based recognition

search module. Its function is to combine and time-align the outputs of the sub-MRNN module and the two RNN modules with the input testing utterance for generating the best recognized base-syllable sequence. Besides, a multilevel pruning method designed to speed up the DP-based recognition process (without degrading the recognition performance) is also proposed [20]. The whole MRNN system can be efficiently trained using a bottom-up hierarchical training scheme. The training algorithm consists of three stages using MCE/GPD algorithms first optimizing the sub-syllable-level recognition, then the base-syllable-level recognition, and lastly, the string-level recognition [13], [21]–[23]. Hence, the four RNN sub-modules are being trained first, then the sub-MRNN, and finally, the whole MRNN.

Some distinct merits of the proposed MRNN-based method shall now be discussed. They include the following.

- This method uses prior knowledge of the phonetic structure of Mandarin speech in the MRNN architecture design and hence provides an interpretable and tractable way to analyze the internal operations of the MRNN, rather than simply treating it as a black box. All constituent RNN modules have their own phonetic meanings.
- This method uses multiple expert RNN modules, therefore, not only taking care of the discrimination of acoustic units but also extracting useful information to enhance the discrimination ability of the recognizer. This information includes the dynamic weighting functions of two broad-class sets (to be discussed in Section II), the syllable-boundary timing information, and the discrimination of intersyllable segments.
- Timing information representing the temporal structure of Mandarin speech signals is directly extracted from observation vectors and used in the recognition search to help solve the time-alignment problem. Duration models, such as state transition probabilities and state duration probability models [2], [24] are therefore not needed.

The proposed MRNN-based method differs from the conventional HMM/ANN hybrid approach in the following two ways. First, the outputs of all RNNs are directly combined to form discriminant functions for speech recognition without being converted to likelihood functions [25], [26]. Second, the proposed method uses ANNs not only to discriminate acoustic units but also to extract useful information in order to enhance the discrimination capability of the recognizer.

The organization of this paper is as follows. Section II presents the proposed MRNN-based CMSR method. The three-stage training method is discussed in Section III. The effectiveness of the proposed method is examined with simulations discussed in Section IV. Some conclusions are given in the last section.

II. Proposed Method

Mandarin Chinese is a tonal and syllabic language. There exists more than 80 000 words, each composed of up to several characters. There are more than 10 000 commonly used characters, each pronounced as monosyllable with one of five tones. There are in total 411 base-syllables, each of which can have up to five different syllabic tones [27], [28]. A complete CMSR system is generally composed of two components: Acoustic processing for syllable identification and Lexical decoding for word (or character) string recognition. In this study, we only consider acoustic processing.

Fig. 1 shows a block diagram of the MRNN recognizer. It consists of four modules: a base-syllable discrimination sub-MRNN module, a boundary-detection RNN module, an intersyllable-segment discrimination RNN module, and a DP-based recognition search module.

The sub-MRNN is an extended version of the MRNN proposed previously for isolated Mandarin base-syllable recognition [13]. Its task is to discriminate 412 base-syllables including a null for silence. It tackles the task by further dividing into four subtasks and deals with each by using a separate RNN submodule. The four subtasks include two which discriminate two sets of subsyllable units and two which generate dynamic weighting functions for two broad-class sets of these recognition units. These two basic recognition unit sets contain 100 right-*final*-dependent (RFD) *initial*s and 39 context-independent (CI) *final*s, respectively. The two broad-class sets of recognition units include one containing three broad-classes of *initial*, *final*, and silence and another containing nine *initial* subclasses divided according to the manner of articulation. Outputs of the two discrimination RNN sub-modules are weighted by these 12 dynamic weighting functions and combined to form the discriminant functions of 412 base-syllables. The way in which dynamic weighting functions are used to form weighted discrimination functions can be regarded as a sophisticated realization of the idea of using weighted distortion sequences [29], [30] or weighted state-likelihoods [15], [31] to improve the performance of speech recognizers. These weighted discrimination functions will therefore possess better discrimination capability.

The task of the boundary-detection RNN module is to detect syllable boundary information to be used in the DP-based recognition search module in order to help solve the time-alignment problem. It uses an RNN to discriminate between syllable boundary and nonsyllable boundary segments.

The function of the intersyllable-segment discrimination RNN module is to generate the discriminant functions of 143 intersyllable diphone-like units to compensate the inter-syllable coarticulation effect. Obviously, this way of handling the coarticulation effect is different from that of the context-dependent HMMs [14]–[18], [28]. In the past, several approaches [19], [32], [33] using similar ideas to explicitly model transitional acoustic units so as to improve speech recognizers were proposed. In [32], syllable boundary information was detected by an ANN and integrated into continuous English speech recognition. It resulted in a reduction of the word error rate by 10%. In [19], inter-syllable segments were statistically modeled and integrated into an HMM-based continuous Mandarin speech recognizer to improve its performance.

The function of the DP-based recognition search module is to integrate the other three main parts and time-align their outputs with the input testing utterance so as to get the best recognized base-syllable sequence. It uses a sophisticated delay-decision, frame-synchronized Viterbi search algorithm to accomplish the job. The DP-based recognition search module

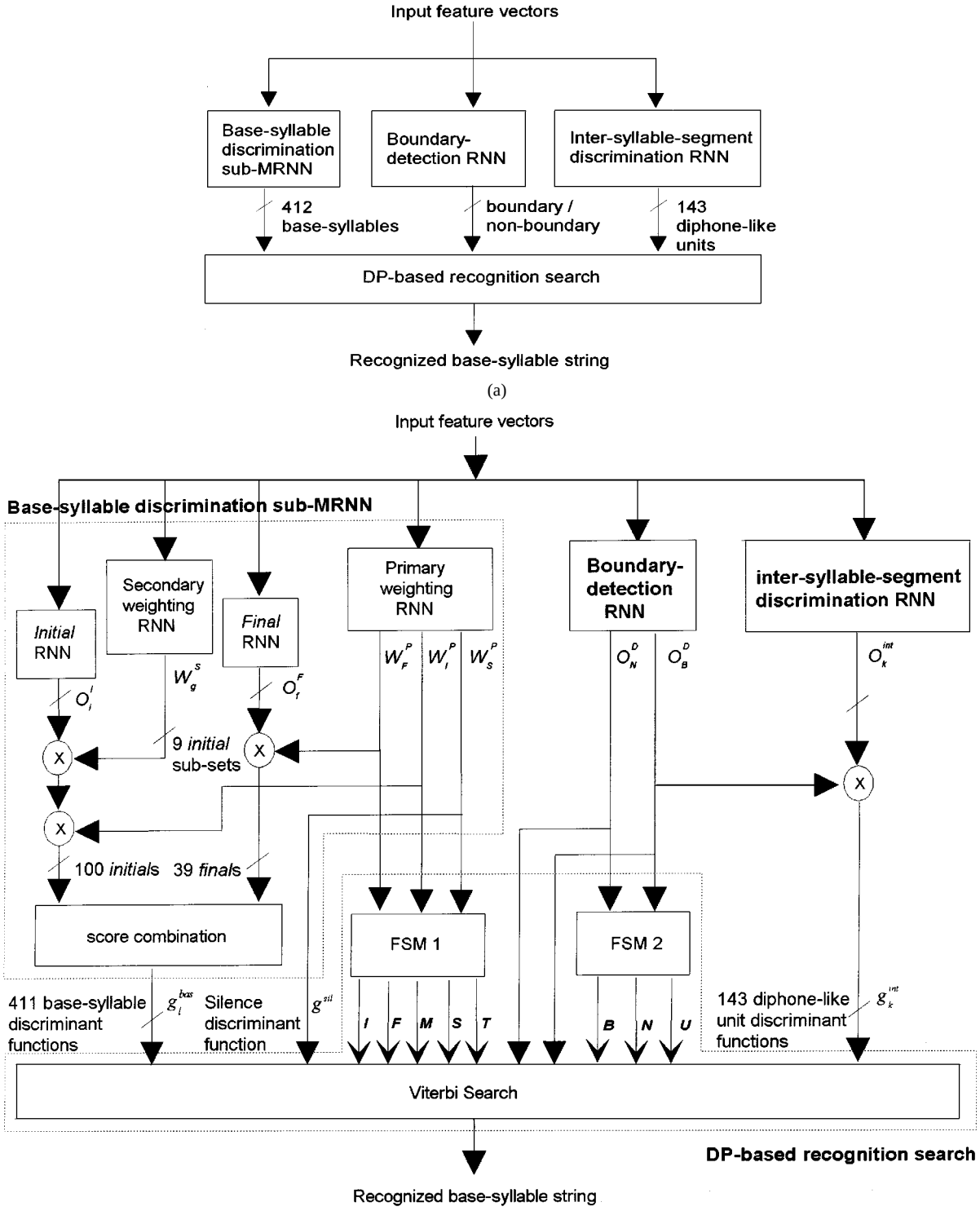


Fig. 1. Proposed MRNN recognizer: (a) schematic diagram and (b) detailed block diagram.

can be further extended by incorporating a multi-level pruning scheme to improve its speed. The whole MRNN can be trained using a bottom-up hierarchical training scheme which is composed of three training stages using MCE/GPD algorithms sequentially optimizing the subsyllable-, base-syllable-, and string-level recognitions. Hence, the four RNN submodules are being trained first, then the sub-MRNN, and finally, the whole MRNN. In the following subsections, we discuss these four main modules and the multilevel pruning scheme in more detail. The bottom-up hierarchical training scheme is discussed in the next section.

A. Base-Syllable Discrimination Sub-MRNN Module

The base-syllable discrimination sub-MRNN module is composed of four RNN submodules. Its design complies with the simple *initial-final* structure of Mandarin base-syllables shown

TABLE I

PHONETIC STRUCTURE AND SUB-SYLLABLE INVENTORY OF MANDARIN BASE-SYLLABLE. (A) THE PHONETIC STRUCTURE OF MANDARIN BASE-SYLLABLE. HERE THE NUMERIC x IN THE TABLE MEANS THE NUMBER OF BASE-SYLLABLES, *INITIALS*, *FINALS*, . . ., ETC. (B) THE SET OF 22 *INITIALS* AND THE NINE *INITIAL* SUB-GROUPS DIVIDED ACCORDING THE MANNER OF ARTICULATION. HERE ϕ_I DENOTES A NULL *INITIAL*. (C) THE SET OF 39 *FINALS*. HERE ϕ_F DENOTES A NULL *FINAL*

Base-syllable 411			
(*initial*) 21	*final* 39		
(consonant) 21	(medial) 3	vowel nucleus 17	(nasal ending) 2

(a)

Initial

Liquid	Nasal	Stop		Affricate		Fricative		Null
		Voiced	Unvoiced	Voiced	Unvoiced	Voiced	Unvoiced	
	m	b	p					ϕ_I
							f	
				tz	ts		s	
l	n	d	t					
				j	ch	r	sh	
				ji	chi		shi	
		g	k				h	

(b)

Final

	Vowel						Di-vowel				Vowel ending with Nasal			
Medial	-ϕ_F	-a	-o	-e	-eh	-er	-ai	-ei	-au	-ou	-an	-en	-ang	-eng
	ϕ_F	a	o	e	eh	er	ai	ei	au	ou	an	en	ang	eng
i-	i	i-a	i-o		i-eh		i-ai		i-au	i-ou	i-an	i-en	i-ang	u-eng
u-	u	u-a	u-o				u-ai	u-ei			u-an	u-en	u-ang	u-eng
iu-	iu				iu-eh						iu-an	iu-en		iu-eng

(c)

in Table I. It can be seen from Table I that Mandarin base-syllables have a very regular, hierarchical phonetic structure which contains two parts: an optional *initial* and a *final*. The *initial* part contains a single consonant if it exists. The *final* part consists of an optional *medial*, a *vowel nucleus*, and an optional *nasal ending*. There are, in total, only 411 base-syllables formed by all legal combinations of 21 *initials* and 39 *finals*. These 39 *finals* are, in turn, formed by the combinations of 3 *medials*, 17 *vowel nuclei*, and 2 *nasal endings*. Due to this simple and regular phonetic structure, these 411 base-syllables form many highly confusable subsets in which base-syllables can only be differentiated by their *initial* consonants, by their short *medials*, or by their *nasal endings* [27], [28]. To efficiently discriminate these 411 highly confusable base-syllables, the sub-MRNN uses the above-mentioned *a priori* knowledge in its architecture design to divide the task into four subtasks and employ the four parallel RNN sub-modules to deal with them separately. The partition into subtasks is performed in both time and feature domains. This design is advantageous in letting each RNN submodule deal with only a part of speech segments and hence ensuring its success in discriminating specific speech patterns.

Specifically, two subtasks are first appointed for the discrimination of two types of speech patterns, *initials* and *finals*, which are separated in time domain with a partial overlap to take into account the intrasyllable coarticulation effect [34].

One RNN submodule is used to tackle the *initial* discrimination subtask and generate discriminant functions for all 100 RFD *initials*, using the *initial* parts of syllables in the input testing utterance. These 100 RFD *initisals* are obtained by expanding the set of 22 CI *initials* using the seven sub-groups of 39 succeeding *finals* divided according to their leading phonemes. Table II shows these seven sub-groups of *finals*.

The use of RFD *initial*s is motivated by the fact that it resulted in better performances in many *initial-final* based HMM recognizers proposed recently [18], [27], [28]. Similarly, another RNN sub-module is used to tackle the *final* discrimination subtask and generate discriminant functions for all 39 CI *finals* using the *final* parts of syllables in the input testing utterance.

Two other subtasks concern the extraction of acoustic cues for the integration of subtasks. They are tackled by two other RNN submodules. One generates primary dynamic weighting functions for the three broad-classes: *initial*, *final*, and silence. The uses of these three primary dynamic weighting functions are three-fold. First, the weighting function for silence is directly taken as the discriminant function of the silence class. Second, the two weighting functions for *initial* and *final* are used

TABLE II
SEVEN SUB-GROUPS OF THE 39 *FINALS* PARTITIONED BY THEIR LEADING PHONEMES FOR DETERMINING THE 1*INITIALS*

Sub-group		*Final*
1	ϕ_F-	ϕ_F
2	a-	a, ai, au, an, ang
3	o-	o, ou
4	e-	e, eh, ei, en, eng, er
5	i-	i, i-a, i-eh, i-ai, i-au, i-ou, i-an, i-en, i-ang, i-eng, i-er
6	u-	u, u-a, u-o, u-ai, u-ei, u-an, u-en, u-ang, u-eng
7	iu-	iu, iu-eh, iu-an, iu-en, iu-eng

to give different weights to the two types of recognition units of *initials* and *finals* to combine the discriminant functions generated by the two subsyllable discrimination RNN sub-modules. Third, they are all used in the multilevel pruning scheme, to help identify stable *initial*-, *final*-, and silence-segments of the input testing utterance in order to set more restricted path constraints to prune unnecessary path searches. The other RNN submodule generates secondary weighting functions for nine broad-classes of *initials*. The main uses of these secondary weighting functions are to provide different weights to the nine groups of *initial* recognition units and help combine their discriminant functions. This has been proved to increase the discrimination capability of the sub-MRNN for distinguishing these 100 RFD *initials* [15], [29], [30].

All four RNNs have the same three-layer structure with all outputs of the hidden layer being fed-back to the hidden layer itself as additional inputs [35]. They all use the same inputs consisting of all recognition feature vectors contained in a window of five frames around the current frame. Their outputs are directly combined to form the discriminant functions of all 411 base-syllables by

$$g_l^{bas}(X;\Lambda_0)(t) = W_I^P(t)\cdot W_g^S(t)\cdot O_i^I(t) + W_F^P(t)\cdot O_f^F(t), \quad l = 1 \sim 411 \quad (1)$$

where

- X — input feature vector sequence of the testing utterance;
- Λ_0 — set of system parameters of the sub-MRNN;
- $W_I^P(t)$ and $W_F^P(t)$ — primary weighting functions for the two broad classes of *initial* and *final*, respectively;
- $O_i^I(t)$ and $O_f^F(t)$ — ith and the fth discriminant functions generated by the *initial* — and *final* — discrimination RNNs for the l-th base-syllable, respectively;
- $W_g^S(t)$ — secondary weighting function of the gth *initial* broad class which contains the ith *initial*.

As for the discriminant function of the silence class, it is directly taken from the primary weighting function for silence, i.e.,

$$g^{sil}(X;\Lambda_0)(t) = W_S^P(t). \quad (2)$$

The validity of using discriminant functions as in (1) can be justified by the superiority of three speech recognition schemes that use similar ideas to improve the performance of speech recognizers. One is the "discriminative weighting distortion sequences" scheme, which enhances the discrimination abilities of isolated speech recognizers by properly weighting their distortion sequences [15], [30]. A speech recognition test showed that the recognition rate for the highly confusable English E-set was increased from 67.6% to 78.1% [30]. A second is the state-weighted HMM method which improves the performance of speech recognizers by discriminatively weighting the state-likelihoods [31]. Our method differs from the state-weighted HMMs by using dynamic weights which depend on the input data instead of static ones which depend on models. A third is the Meta-Pi network in which several ANN modules are first used to estimate the probabilities of the input pattern generated by different sources (e.g., speakers) [10]. Their outputs are then combined to form the *a posterior* probabilities for robust multisource recognition. Our method differs from the Meta-Pi network because it uses the base-syllable-level MCE/GPD training algorithm to design the combination function.

B. Boundary-Detection RNN Module

The boundary-detection RNN module uses a single RNN to generate two dynamic timing functions, $O_B^D(t)$ and $O_N^D(t)$, for syllable boundary and nonsyllable boundary segments, respectively. The RNN has the same structure as the four RNNs used in the aforementioned sub-MRNN. The input features include seven observation vectors contained in a window of the current frame. The main use of these two dynamic timing functions is to combine the discriminant functions of 412 base-syllables and of 143 intersyllable diphone-like units to form the discriminant function for each candidate base-syllable string. A similar idea was used in [32] which used syllable boundary information, detected by an ANN, to improve continuous English speech recognition. It is worthwhile to note that these two timing functions contribute in a different way to the recognition search compared with the conventional HMM method. They provide explicit timing cues of the input testing utterance and give scores directly to assert or object to all base-syllable-to-base-syllable transitions in the recognition search; while the HMM method provides implicit temporal constraints to the recognition search by giving state transition probabilities, using state duration probability models [1], [24], or setting state duration bounds [36]. Obviously, using these two timing functions is also different from the constant duration-penalty scheme [2] used by a recognizer to suppress the insertion errors. Besides, these two timing functions are also used in the multilevel pruning scheme to suppress unnecessary searches for base-syllable-to-base-syllable transitions.

The RNN can be trained by assigning timing functions related to syllable boundaries as output targets. Ideally, we can use a single-frame impulse as the target timing function of $O_B^D(t)$ to precisely locate a base-syllable boundary. However, in practical implementation, it is difficult to model such a stringent timing function using the RNN technique. We therefore relax the requirement by training the RNN using an output target of three-frame pulse for each base-syllable boundary.

TABLE III
(a) TWELVE RIGHT-*INITIAL* CLASSES AND (b) TWELVE LEFT-*FINAL*CLASSES FOR DETERMINING THE SET OF 143 INTERSYLLABLE DIPHONE-LIKE UNITS

Sub-group		Right *initial* context
1	Stop-	b, d, g, p, t, k
2	Affricate-	tz, j, ji, ts, ch, chi
3	Fricative-	f, s, sh, shi, h
4	r-	r
5	n-	n
6	l-	l
7	m-	m
8	ϕ_I-	a, ai, au, an, ang, o, ou, e, eh, ei, en, eng, er
9	i-	i, i-a, i-o, i-eh, i-ai, i-au, i-ou, i-an, i-en, i-ang, i-eng
10	u-	u, u-a, u-o, u-ai, u-ei, u-an, u-en, u-ang, u-eng
11	iu-	iu, iu-eh, iu-an, iu-en, iu-eng
12	Silence-	Silence

(a)

Sub-group		Left *final* context
1	-ϕ_F	ϕ_F
2	-a	a, i-a, u-a
3	-o	o, u-o, i-o
4	-e	e
5	-i	ai, ei, i, i-ai, u-ai, u-ei
6	-eh	eh, i-eh, iu-eh
7	-u	u, au, ou, iu, i-au, i-ou
8	-an	an, en, i-an, i-en, u-an, u-en, iu-an, iu-en
9	-ang	ang, eng, i-ang, i-eng, u-eng, u-ang
10	-iu	iu
11	-er	er
12	-Silence	Silence

(b)

C. Inter-Syllable-Segment Discrimination RNN Module

The function of the inter-syllable-segment discrimination RNN is to generate the discriminant functions, $O_k^{int}(t)$, for 143 inter-syllable diphone-like units formed by all meaningful combinations of 12 left-*final* classes (including silence) and 12 right-*initial* classes (including silence). These two sets are shown in Table III [15]. The 12 left-*final* classes are formed by partitioning 39 *finals* according to their ending phonemes. The 12 right-*initial* classes are formed by partitioning the generalized *initial* set, composing of 21 *initials* and 38 null-*initial*s (i.e., 38 base-syllables with null *initial*), according to the manner of articulation with sonorants being specially considered. Specifically, they include *stop*, *affricate*, *fricative*, {/r/}, {/n/}, {/l/}, {/m/}, four null-*initial* classes, and silence. The four null-*initial* classes are formed by partitioning all 38 null-*initial*s according to the 4 leading *medial*s including /i/, /u/, /iu/, and a null. The approach of determining the set of inter-syllable diphone-like units is a tradeoff between the importance of unit on the inter-syllable coarticulation effect and the total number of units. The inter-syllable-segment discrimination RNN has the same three-layer structure with input features including seven observation vectors contained in a window of the current frame. The RNN is trained using speech segments located around all syllable boundaries in the training set. For each syllable boundary, output targets of six frames centered around it are set for training.

It is worthwhile to note that the MRNN system handles the coarticulation effect in a way different from the HMM method using context-dependent models [14]–[18], [28]. It takes a pattern recognition approach to directly classifying speech segments carrying intersyllable coarticulations, while the context-dependent HMMs try to split models of acoustic units according to the context. Recently, several approaches using similar ideas to model transitional acoustic units for improving the performance of HMM-based speech recognizers were proposed [32], [33].

D. DP-Based Recognition Search Module

The function of the DP-based recognition search module is to combine and time-align the outputs of the other three main modules with the input testing utterance for generating the best recognized base-syllable sequence. The discriminant function for a path Q of a candidate base-syllable sequence S is defined by

$$g^{str}(X,S,Q;\Lambda)=\frac{1}{L}\sum_{t=0}^{L-1}\left\{g^{syl}_{q_{syl}(t)}(t)+g^{tim}_{q_{tim}(t)}(t)+g^{int}_{q_{int}(t)}(t)\right\} \tag{3}$$

where L is the length of the input feature vectors, $Q=\{(q_{syl}(t),q_{tim}(t),q_{int}(t)),\ t=0,1,\ldots,L-1\}$ is the legal path for S, Λ is the set of system parameters for the MRNN, and $q_{syl}(t)$ is the base-syllable state of the path Q at t

$$q_{tim}(t)=\begin{cases}1, & \text{if } q_{syl}(t)\neq q_{syl}(t-1)\\ 0, & \text{if } q_{syl}(t)=q_{syl}(t-1)\end{cases} \tag{4}$$

is the timing (i.e., transition/nontransition) state of the path Q at t

$$q_{int}(t)=\begin{cases}k, & \text{if } q_{syl}(t)\neq q_{syl}(t-1)\\ 0, & \text{if } q_{syl}(t)=q_{syl}(t-1)\end{cases} \tag{5}$$

is the intersyllable unit state of the path Q at t with k determined by $q_{syl}(t)$ and $q_{syl}(t-1)$

$$g^{syl}_{q_{syl}(t)}(t)=\begin{cases}g^{bas}_l(X;\Lambda_0)(t), & \text{if } q_{syl}(t) \text{ is the } l\text{-th base-syllable}\\ C_S\cdot g^{sil}(X;\Lambda_0)(t), & \text{if } q_{syl}(t) \text{ is a silence}\end{cases} \tag{6}$$

is the base-syllable discriminant score at t

$$g^{tim}_{q_{tim}(t)}(t)=\begin{cases}C_D\cdot O^D_B(t) & \text{if } q_{tim}(t)=1\\ C_D\cdot O^D_N(t) & \text{if } q_{tim}(t)=0\end{cases} \tag{7}$$

is the timing score at t

$$g^{int}_{q_{int}(t)}(t)=\begin{cases}C_{int}\cdot O^D_B(t)\cdot \overline{O}^{int}_k(t) & \text{if } q_{int}(t)=k\\ 0 & \text{otherwise}\end{cases} \tag{8}$$

is the intersyllable-segment discriminant score at t

$$\overline{O}^{int}_k(t)=\frac{1}{6}\sum_{\tau=t-3}^{t+2}O^{int}_k(\tau) \tag{9}$$

is the discriminant function for the kth inter-syllable diphone-like unit at t; and C_S, C_D, and C_{int} are weighting constants. The final goal of the recognition search is to find

out the best base-syllable sequence $\hat{S}$ among all possible sequences. The decision rule is thus defined by

$$\hat{S} = \arg\max_{S} \max_{S,Q} g^{str}(X, S, Q; \Lambda). \tag{10}$$

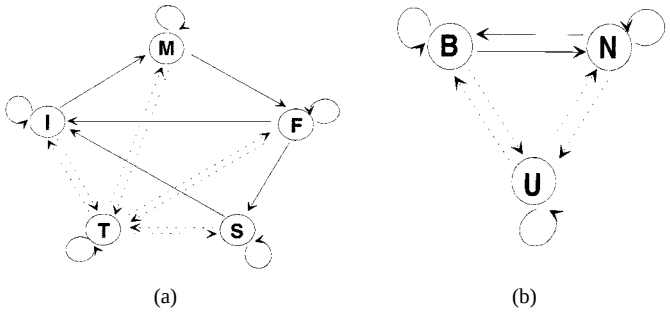
It is noted that in the above formulation, each base-syllable is represented by a single state.

Based on the aforementioned criterion, a modified one-stage Viterbi search algorithm is proposed to find out the best base-syllable sequence. It retains multiple cumulative scores for each base-syllable by expanding its single-state structure into a series of clone states. This modification is designed to allow the Viterbi search algorithm to delay the decision making for base-syllable-to-base-syllable transition until the end of a finite-duration pulse of $O_B^D(t)$; otherwise, it will always occur at the beginning frame of the pulse if the weighting constant C_D is large. This maintains fair competition for all candidate frames within the pulse to become a true base-syllable-to-base-syllable transition. The length of the delay can be a value larger than the maximum pulse-width of $O_B^D(t)$. Aside from using the timing function generated by the boundary-detection RNN module, the Viterbi search algorithm also uses the path constraints of base-syllable duration bounds to eliminate insertion errors with a very short base-syllable duration.

E. Multilevel Pruning Scheme

The multilevel pruning scheme is designed to be incorporated into the DP-based recognition search to improve its speed. The design goal is to maximize the reduction of unnecessary path searches while maintaining recognition performance and consuming minimum extra efforts. It takes useful acoustic cues directly from four RNNs of the MRNN to accomplish its job. It is a combination of the following three pruning schemes: base-syllable deactivation, preclassification-based pruning and presegmentation-based pruning. The base-syllable deactivation scheme uses an idea similar to the phone deactivation method [37]. It eliminates the searches for all paths containing unlikely base-syllables with very low frame-based discriminant functions. The preclassification based pruning scheme uses the idea of setting more restricted search constraints for the stable parts of the input testing utterance to eliminate unnecessary path searches [20]. The presegmentation based pruning scheme uses a similar idea to set more restricted inter-base-syllable transition/nontransition constraints for predetected boundary and nonboundary parts of the input testing utterance to eliminate unnecessary interbase-syllable transition/nontransition tests in the recognition search. We will discuss the latter two schemes in more detail.

As mentioned previously, the primary weighting RNN generates three outputs to discriminate each input frame among the three broad classes of silence, *initial*, and *final*. The preclassification based pruning scheme uses these three outputs to drive a preclassification finite state machine (FSM) [see Fig. 2(a))] to classify and label each input frame into four stable states of silence (***S***), *initial* (***I***), *medial* (***M***), and *final* (***F***), and one transient (***T***) state. Specifically, the pre-classification FSM compares the three outputs of the primary weighting RNN with two threshold values, TH_L and TH_H. While one (i.e., *initial*, *final* or silence) output is higher than TH_H and the other two are all lower than TH_L, the FSM moves into the corresponding stable (***I, F***, or ***S***) state if it is a legal one. When both the *initial* and *final* outputs are higher than TH_H and the silence output is lower than TH_L, the FSM moves into the ***M*** state. Otherwise, it goes to the ***T*** state. Similarly, the presegmentation based pruning scheme uses a presegmentation FSM [see Fig. 2(b)], driven by the difference between the two outputs of the boundary-detection RNN (i.e., $O_B^D(t) - O_N^D(t)$) to classify and label each input frame into two certain states of boundary (***B***) and nonboundary (***N***), and one uncertain (***U***) state.



(a) (b)

Fig. 2. State transition diagrams of (a) the preclassification FSM and (b) the presegmentation FSM.

We then use the output labels of these two FSMs to explicitly model the temporal structure of the input speech signal. Those FSM labels are incorporated into the DP-based recognition search module, so we can prune unnecessary path searches. Each base-syllable is represented by three left-to-right states: *initial*, *medial* and *final*. When an input frame is labeled ***I***, ***M,*** or ***F*** state, we only allow the frame to stay in the corresponding *initial*, *medial*, or *final* states of all base-syllables. If it is labeled as an ***S*** state, we let it stay in the silence state. If it is labeled an ***N*** state, no base-syllable-to-base-syllable transitions are allowed to occur at this frame. For B states, we make base-syllable-to-base-syllable transition decisions and allow one and only one transition to occur in a segment of consecutive B states. Lastly, for both ***T*** and ***U*** states, a full search is performed. It is worthwhile to note that, from the viewpoint of the DP search, the resulting path constraining scheme is a partial-hard-decision-and-partial-soft-decision one. Besides, this pruning scheme can be used in conjunction with some path pruning techniques, such as the beam search, to further improve the recognition efficiency [20].

III. Bottom-Up Hierarchical Training Scheme

To efficiently train the MRNN recognizer, a three-stage training method sequentially applying the subsyllable-, base-syllable-, and string-level MCE/GPD algorithms [14],[21]–[23] is proposed. In the first stage, the two RNNs for discriminating the two sub-syllable sets of *initial* and *final* are trained separately using the subsyllable-level MCE/GPD algorithm. Meanwhile, the primary and secondary weighting RNNs are trained accordingly using the backpropagation through time algorithm with "0–1" output target functions [38]. Here, all targets are set based on an HMM-based segmentation of the training set. The four RNNs are then combined together in the second training stage to form the base-syllable discrimination sub-MRNN and

fine-tuned by the base-syllable-level MCE/GPD algorithm. The training goal is to make the sub-MRNN be a good classifier for all 412 base-syllables. In the third training stage, the boundary-detection RNN and the intersyllable-segment discrimination RNN are first separately trained by the backpropagation through time algorithm with "0–1" output target functions and the MCE/GPD algorithm, respectively. They are then integrated with the base-syllable discrimination sub-MRNN to form the MRNN recognizer. The whole MRNN is finally fine-tuned by the string-level MCE/GPD algorithm.

It is worthwhile to note that the "0–1" bounds of all primary and secondary dynamic weighting functions, set as learning targets in the first-stage training, are relaxed and freed of any manual control in the following training stage. The ultimate levels that these weighting functions may reach are therefore determined automatically. The two dynamic timing functions are treated similarly. This target-setting scheme is advantageous in enhancing the discrimination capacity of the MRNN recognizer via automatically putting special emphases on the most distinguishing parts of the input testing utterance.

IV. Simulations

The effectiveness of the proposed method was examined with simulations using a continuous Mandarin speech database uttered by a single male speaker. The database contains 452 sentential utterances and 200 paragraphic utterances. Texts of these 452 sentential utterances are well-designed, phonetically balanced sentences. Texts of these 200 paragraphic utterances are news selected from a large news corpus covering a variety of subjects. There are, in total, 6021 and 29 073 syllables in the sentential- and paragraphic-utterance sets, respectively. All utterances were spoken naturally at a speed of about 4.5 syllables/s. The database was divided into two parts. The one containing 28 060 base-syllables with a length of about 3.5 h was used for training and the other containing 7034 base-syllables with a length of about half an hour was used for testing. All speech signals were A/D-converted at a rate of 10 kHz and preemphasized with a digital filter, $1-0.95z^{-1}$. They were then analyzed to extract recognition features for each 20-ms Hamming-windowed frame with a 10-ms frame shift. The recognition features included 12 MFCCs, 12 delta MFCCs, 12 delta-delta MFCCs, a delta log-energy, and a delta-delta log-energy. All RNNs used in the MRNN recognizer had the same three-layer, simple recurrent structure with all outputs of the hidden layer being fed back to the hidden layer itself as additional inputs [35]. The length of the input window was seven frames for both the boundary-detection RNN and the intersyllable-segment discrimination RNN, and was five frames for the other four RNNs. All output-layer nodes of the six RNNs used linear output functions instead of nonlinear *sigmoid* functions to make their responses have larger dynamic ranges.

First, we checked the performance of the MRNN system. Two versions of the system were realized and tested. The first one was a basic recognizer without invoking the intersyllable-segment discrimination RNN module. It is referred to as MRNN-B. The second one, referred to as MRNN-I, was the complete system with the intersyllable-segment discrimination RNN module being included. In the test, eight delayed clone states for each base-syllable were sustained in the Viterbi search algorithm to keep the cumulative discriminant scores. The number of best base-syllable strings involved in the string-level MCE/GPD algorithm was empirically set to 20. The recognition results of the two versions are displayed in Fig. 3. The best base-syllable accuracy rates were 86.8% and 87.3% for MRNN-B and MRNN-I, respectively.

For performance comparison, the HMM method was also tested. To determine the basic recognition units, two popular approaches used in the state-of-the-art HMM-based continuous Mandarin speech recognition systems [14]–[18], [28] proposed in recent years were considered. One approach uses context-dependent (CD) phone-like units and another uses right-context-dependent (RCD) *initial* and *final* units [15]–[18]. According to the experimental results of the two comparative studies done in [16] and [18], these two approaches performed almost equally well in both speaker-independent (SI) and speaker-dependent (SD) cases. Besides, many other studies also confirmed that the right-context dependency is more important than the left-context dependency in both approaches of using phone-like units and *initial-final* units [14], [17], [18]. Based on the above conclusions, the second approach of using RCD *initial* and *final* units was chosen in the test. To compare with the MRNN-B recognizer, a baseline system using 100 RFD *initial*s and 39 CI *final*s as basic recognition units was first built. It is referred to as the HMM-B recognizer. A three-state HMM model was trained for each of these 100 RCD *initial* unit. A five-state model was trained for each *final* unit. For silence, a single-state model was used. The number of mixture Gaussian components in each state of these HMM models varied from one to M depending on the amount of training data. All HMM models were ML-trained and tested using the same database mentioned previously. The experimental results are also displayed in Fig. 3. The best base-syllable accuracy rate was 82.9%. Obviously, this result was inferior to that achieved by the MRNN-B recognizer. It can also been seen from Fig. 3 that the number of parameters used in the best ML-trained HMM-B recognizer is about two times of that used in the best MRNN-B recognizer. So the ML-trained HMM-B recognizer has higher computational complexity.

For a fairer comparison, the baseline HMM-B recognizer was further refined using a string-level MCE/GPD training algorithm. In the training, the MCE/GPD algorithm used top 20 best base-syllable sequences to fine-tune the HMM models. The experimental results are also displayed in Fig. 3. The best base-syllable accuracy rate increased to 86.0%. This performance is comparable to that of the MRNN-B recognizer. As for the system complexity, the best MCE/GPD-trained HMM-B recognizer used less parameters as compared with the best MRNN-B recognizer.

To compare with the MRNN-I recognizer, the baseline HMM-B recognizer was extended to become an RCD HMM system by letting all *final* models right depend on the *initial* of the succeeding base-syllable. It is referred to as the HMM-RCD recognizer. It used a decision tree to select appropriate contextual models for each *final*. The expanded tree shown in Fig. 4 was built using the phonetic knowledge of Mandarin base-syllables. A rule based on comparing the number of training

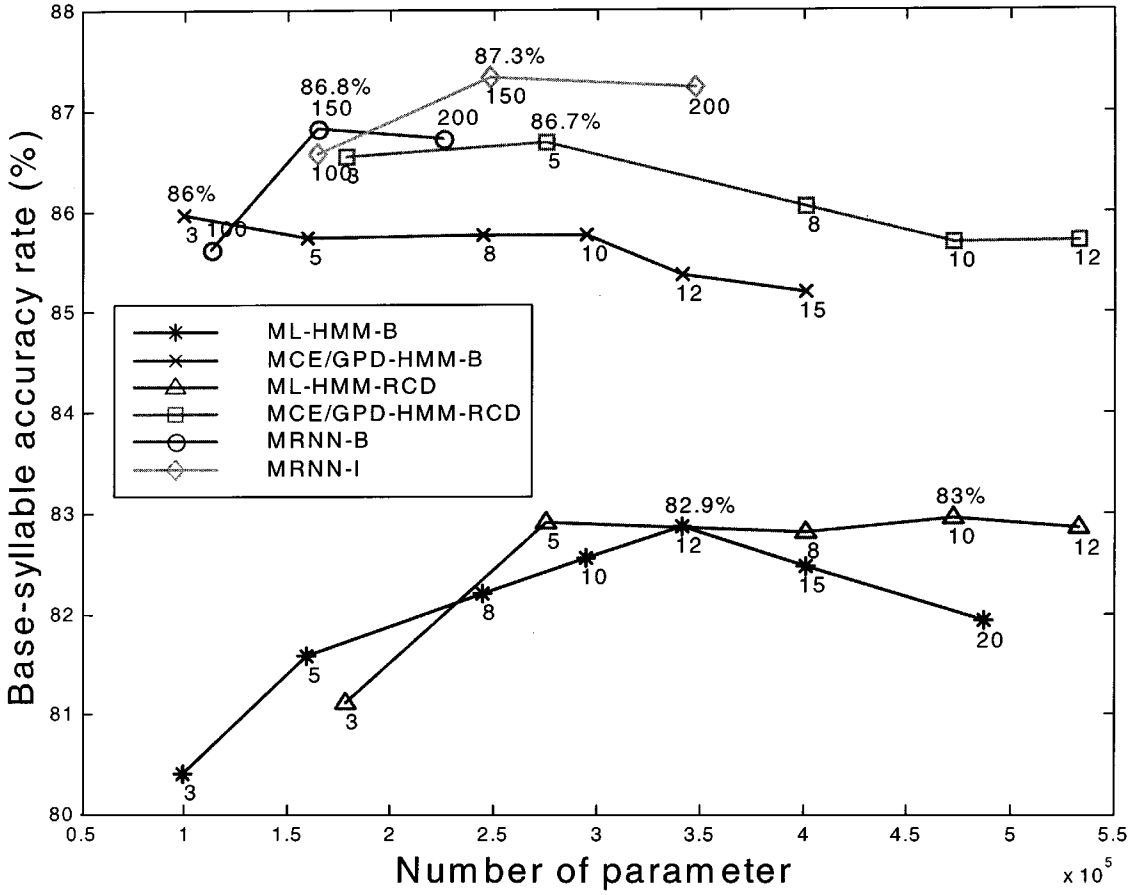


Fig. 3. Experimental results for two MRNN-based recognizers, MRNN-B and MRNN-I, and four HMM-based recognizers, ML-HMM-B, ML-HMM-RCD, MCE/GPD-HMM-B and MCE/GPD-HMM-RCD. Here the x ($x = 3, 5, 8, 10, 12, 15, 20$) for HMM recognizers denotes the maximum mixture number used in each state and the y ($y = 100, 150, 200$) for MRNN recognizers denotes the number of the hidden neurons used in the *initial*, *final*, and intersyllable-segment discrimination RNNs.

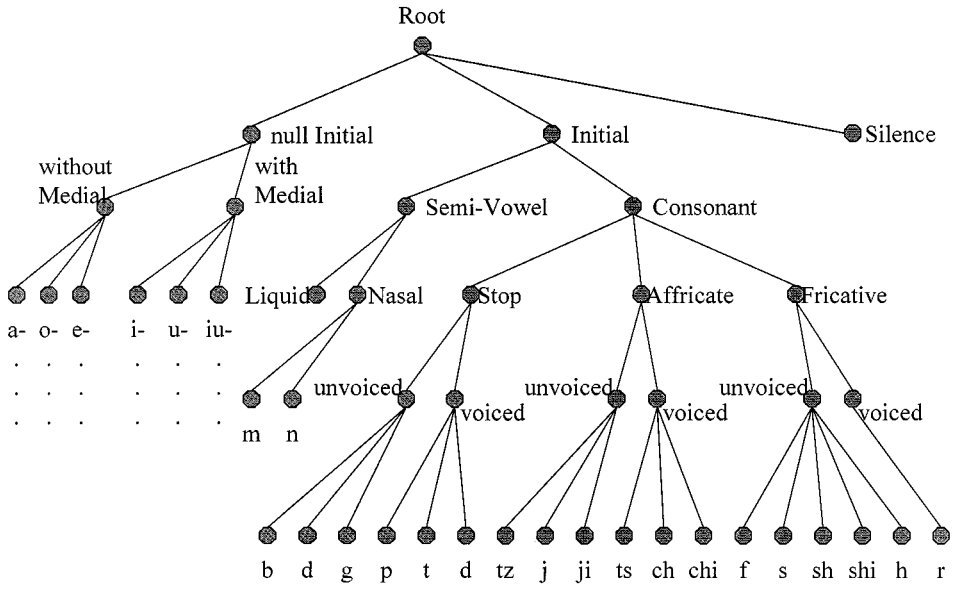


Fig. 4. Decision tree used in the HMM approach to choose the right-context-dependent *final* models.

samples with a pre-determined threshold (50 training samples per model) was applied to determine the clusters of contexts. A total of 210 RCD *final* models were finally constructed. They all used the same five-state HMM structure. All models are trained using the ML criterion. To efficiently use the training data, the first three states of all RCD models of the same *final* were tied together. The experimental results are also displayed in Fig. 3. It can be found from Fig. 3 that the best base-syllable

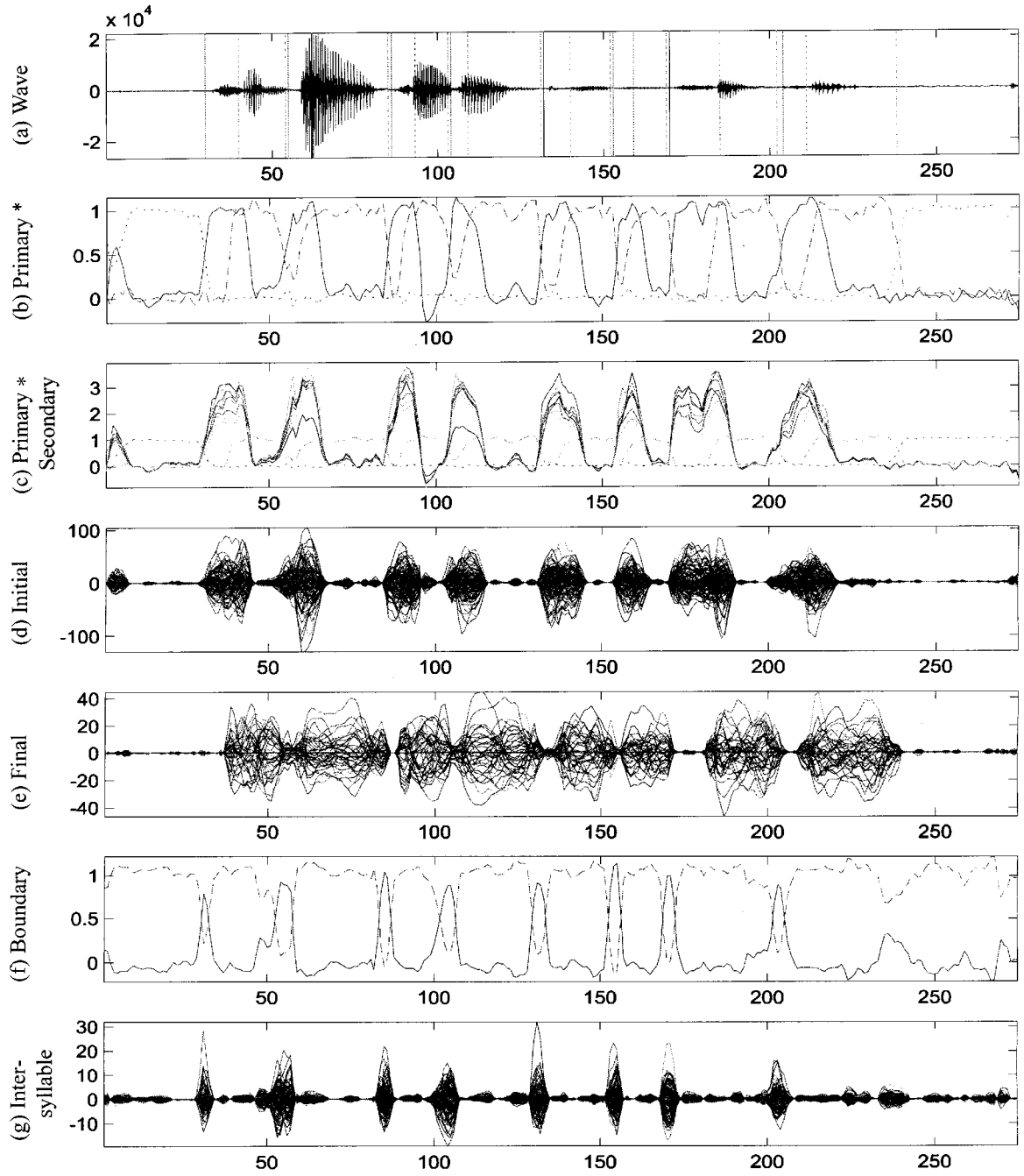


Fig. 5. Typical example showing the responses of the MRNN-I recognizer to the input utterance "/chi-i-eng/ /b-a/ /j-e/ /l-an/ /t-u/ /tz/ /s-u-eng/ /tz-ou/" (vertical lines denotes silence */initial/final* boundaries given by the HMM-based method): (a) input waveform, (b) three outputs of the primary weighting RNN (solid line: *initial*, dashed line: *final*, dotted line: silence), (c) 11 weighting functions added to subsyllable recognition units (solid lines: *initials*, dashed line: *final*, dotted line: silence), (d) weighted discriminant functions for 100 RFD *initials*, (e) weighted discriminant functions for 39 CI *finals*, (f) two outputs of the boundary-detection RNN (solid line: syllable boundary, dashed line: nonsyllable boundary), and (g) weighted discriminant score for 143 intersyllable diphone-like units.

accuracy rate was raised to 83%. This performance is still much lower than the MRNN-I recognizer. It is worth to note that these results are comparable to those obtained in a recent SD Mandarin speech recognition (MSR) study using a female-speaker data set of the HKU93 database [16], and is much inferior to those achieved in other two SD MSR studies [15], [28]. If we consider the fact that both databases used in [28] and [15] are much smaller and contain only short sentential utterances with all syllables being clearly pronounced, the performance differences are reasonable. Finally, the MCE/GPD training algorithm was applied to improve the HMM-RCD recognizer. The best recognition rate is 86.7% which is comparable to that of the MRNN-I recognizer. As for the system complexity, the best HMM-RCD recognizer is higher because it used more parameters than that of the best MRNN-I recognizer. Besides, it needs much larger memory space to perform its recognition search.

We then used the MRNN-I to examine the efficiency of the multilevel pruning scheme. Each base-syllable is now represented by three left-to-right states designated, respectively, as

TABLE IV
EXPERIMENTAL RESULTS OF THE MRNN RECOGNIZER INVOKING WITH THE MULTI-LEVEL PRUNING SCHEME

	Thresholds	Active syllable states	Active syllable transition paths	Base-syllable accuracy rate
Baseline	N/A	100.0%	100.0%	86.9%
Multi-level pruning	TH_D: -50/-30/-30 TH_C: 0.8/0.2 TH_S: 0.25/-0.95	48.6%	23.0%	87.1%
	TH_D: -40/-20/-20 TH_C: 0.775/0.225 TH_S: -0.00/-0.95	45.6%	21.8%	86.8%

initial, *medial*, and *final*. All threshold values of the two FSMs were empirically determined to be as loose as possible while keeping the recognition accuracy almost the same as in the baseline system. In this test, they were set to be $0.8/0.2$ for the preclassification pruning, $0.25/-0.95$ for the presegmentation pruning, and $-50/-30/-30$ for the base-syllable deactivation pruning. Here, the more stringent value of -50 was set for the *initial* state while another value of -30 was set for both the *medial* and *final* states. This setting allowed more *initial* states to survive for the initial parts of the input utterance. The experimental results are listed in Table IV. It can be seen from the table that the performance of the baseline system degraded slightly to 86.9% as we changed the number of clone states for each base-syllable from eight to three. It can also be found from Table IV that only 48.6% of the surviving base-syllable states and 23% of the surviving base-syllable transitions were needed to be considered in the search without degrading the recognition accuracy. This is a moderate saving.

To illustrate the MRNN, a typical example of the responses of the MRNN-I to an input sentential utterance is shown in Fig. 5. This figure reveals information to explain the detailed operations of the MRNN. First, it can be seen from Fig. 5(b) that the primary weighting functions for the three broad classes of *initial*, *final*, and *silence* respond very well, respectively, to the syllable-*initial*, syllable-*final*, and silence parts of the input speech signal. They can therefore provide proper weights to identify the recognition units belonging to the correct broad classes as well as to reject those belonging to the incorrect broad classes for each of the above three parts of the input speech signal. This can be confirmed by examining the weighted discriminant functions for 100 RFD *initials* and 39 CI *finals* displayed in Fig. 5(d) and (e), respectively. Second, it can also be seen from Fig. 5(b) that the active regions of the two primary weighting functions for *initial* and *final* broad classes are overlapped by several frames. This makes the MRNN use the *initial-final* transitional segments of the input speech signal simultaneously in both the *initial* and *final* discriminations. The intrasyllable coarticulation effect can therefore be partially compensated. According to a previous study [34], this is of great help in distinguishing highly confusable Mandarin base-syllables. Third, the actual weighting functions added to the recognition units of the *final* broad-class, silence, and the nine *initial* subclasses are displayed in Fig. 5(c). It can be seen from Fig. 5(c) that all weighting functions for these nine *initial* sub-classes have much higher averaged active levels than those for the *final* broad-class. This shows that the MRNN recognizer relies more heavily on the syllable-*initial* part in the base-syllable discrimination. This complies with our *a priori* linguistic knowledge that *initials* are more important for distinguishing highly confusable base-syllables [15] as in the English E-set recognition [29], [30]. Fourth, as shown in Fig. 5(f), the two timing functions generated by the boundary-detection RNN module respond well to most base-syllable boundaries. They can therefore provide proper timing cues to help solve the time-alignment problem. Lastly, the weighted intersyllable-segment discriminant scores are displayed in Fig. 5(g). By carefully examining these scores, we find that they respond well to the 143 intersyllable diphone-like units.

V. CONCLUSIONS

In this paper, a new MRNN-based method for CMSR has been discussed. It differs from the conventional HMM/ANN hybrid approach by applying prior domain knowledge in the MRNN architecture design to build up multiple RNN modules to extract more useful information, other than the basic discrimination information of acoustic units, to enhance the ability of the recognizer in discriminating the vocabulary of 412 highly confusable Mandarin base-syllables. Experimental results have confirmed that it outperformed the ML-trained HMM method and was comparable to the MCE/GPD-trained HMM method. It is therefore a promising method for CMSR. Further studies to extend the method by incorporating more modules to tackle the tasks of tone recognition and lexical decoding will be performed in the future. It would also be worth considering applying this method to other language.

ACKNOWLEDGMENT

The authors thank Chunghwa Telecommunication Laboratories, R.O.C., for kindly supplying the database.

REFERENCES

[1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.
[3] H. Bourlard and N. Morgan, *Connectionist speech recognition – A hybrid approach*. Norwell, MA: Kluwer, 1994.
[4] N. Morgan and H. Bourlard, "Continuous speech recognition – An introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Mag.*, vol. 12, pp. 25–42, May 1995.
[5] M. Franzini, K. Lee, and A. Waibel, "Connectionist Viterbi training: a new hybrid method for continuous speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1990, pp. 425–428.
[6] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, pp. 298–305, Mar. 1994.
[7] R. A. Jacobs, M. I. Jordan, and A. G. Barto, "Task decomposition through competition in a modular connectionist architecture: the what and where vision tasks," *Cogn. Sci.*, vol. 15, pp. 219–250, 1991.
[8] B. L. M. Happel and J. M. J. Murre, "Design and evolution of modular neural network architectures," *Neural Networks*, vol. 7, no. 6/7, pp. 985–1004, 1994.
[9] A. Waibel, H. Sawai, and K. Shikano, "Modularity and scaling in large phonemic neural networks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1888–1898, Dec. 1989.
[10] J. B. Hampshire II and A. Waibel, "The Meta-Pi network: building distributed knowledge representations for robust multi-source pattern recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 751–769, 1992.
[11] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Comput.*, vol. 6, pp. 181–214, 1994.

[12] J. Fritsch and M. Finke, "ACID/HNN: Clustering hierarchies of neural networks for context-dependent connectionist acoustic modeling," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 3, 1998, pp. 505–508.
[13] S. H. Chen and Y. F. Liao, "'Modular recurrent neural networks for Mandarin syllables recognition'," *IEEE Trans. Neural Networks*, vol. 9, pp. 1430–1441, Nov. 1998.
[14] C. H. Lee and B. H. Juang, "A survey on automatic speech recognition with an illustrative example on continuous speech recognition of Mandarin," *Comput. Linguist. Chinese Lang. Process.*, vol. 1, pp. 1–36, Aug. 1996.
[15] J.-L. Shen, "Continuous Mandarin speech recognition for Chinese language with large vocabulary based on segmental probability model," *IEE Proc. Visual Image Signal Processing*, vol. 145, pp. 309–315, Oct. 1998.
[16] J. J.-X. Wu, L. Deng, and J. Chan, "Modeling context-dependent phonetic units in a continuous speech recognition system for Mandarin Chinese," in *Int. Conf. Spoken Language Processing*, 1996, pp. 2281–2284.
[17] B. Ma, T. Huang, B. Xu, X. Zhang, and F. Qu, "Context-dependent acoustic models for Chinese speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1996, pp. 455–458.
[18] F. Seide and N. J. C. Wang, "Phonetic modeling in the Philips Chinese continuous-speech recognition system," in *Proc. Int. Symp. Chinese Spoken Language Processing*, 1998.
[19] S. Chang and S. H. Chen, "Improved syllable-based continuous Mandarin speech recognition using inter-syllable boundary models," *Electron. Lett.*, vol. 31, pp. 853–854, May 1995.
[20] S. H. Chen, Y. F. Liao, S. M. Chiang, and S. Chang, "An RNN-based preclassification method for fast continuous Mandarin speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 86–89, Jan. 1998.
[21] B. H. Juang, W. Chou, and C. H. Lee, "Statistical and discriminative methods for speech recognition," in *Automatic Speech And Speaker Recognition – Advanced Topics*. Norwell, MA: Kluwer, 1996, pp. 109–132.
[22] ——, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 257–265, May 1997.
[23] J. K. Chen and F. K. Soong, "An N-best candidates-based discriminative training for speech recognition application," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 206–216, Jan. 1994.
[24] D. Burshtein, "Robust parametric modeling of durations in hidden Markov Models," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 240–242, May 1996.
[25] A. Senior and T. Robinson, "Forward-backward retraining of recurrent neural networks," in *Advances Neural Information Processing Systems 8 (NIPS'8)*, 1996, pp. 743–749.
[26] D. Kershaw, T. Robinson, and M. Hochberg, "Context-dependent classes in a hybrid recurrent network-HMM speech recognition system," in *Advances Neural Information Processing Systems 8 (NIPS'8)*, 1996, pp. 750–756.
[27] L. S. Lee, C. Y. Tseng, H. Y. Gu, F. H. Liu, C. H. Chang, Y. H. Lin, and Y. L. etc, "Golden Mandarin (I) — a real-time Mandarin speech dictation machine for Chinese language with very large vocabulary," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 2, pp. 158–179, 1993.
[28] H. M. Wang, T. H. Ho, R. C. Yang, J. L. Shen, B. R. Bai, J. C. Hong, W. P. Chen, T. L. Yu, and L. S. Lee, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 195–200, Mar. 1997.
[29] P. C. Chang and B. H. Juang, "Discriminative training of dynamic programming based speech recognizers," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 2, pp. 135–143, 1993.
[30] P. C. Chang, S. H. Chen, and B. H. Juang, "Discriminative analysis of distortion sequences in speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 3, pp. 326–333, 1993.
[31] O. W. Kwon and C. K. Un, "Discriminative weighting of HMM state-likelihood using the GPD method," *IEEE Signal Processing Letters*, vol. 3, no. 9, pp. 257–259, 1996.
[32] S.-L. Wu, M. L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1997, pp. 987–990.
[33] R. Gemello, D. Albesano, and F. Mana, "Continuous speech recognition with neural networks and stationary-transitional acoustic units," in *In Int. Conf. on Neural Networks*, vol. 4, 1997, pp. 2107–2111.
[34] H. C. Wang and H. F. Pai, "Recognition of Mandarin syllables based on the distribution of two-dimensional cepstral coefficients," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 8, no. 1, pp. 247–257, 1993.
[35] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, pp. 179–211, 1990.
[36] H. Y. Gu, C. Y. Tseng, and L. S. Lee, "Isolated-utterance speech recognition using hidden Markov models with bounded state duration," *IEEE Trans. Signal Processing*, vol. 39, pp. 1743–1752, Aug. 1991.
[37] S. Renals and M. Hochberg, "Efficient search using phone probability estimates," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1995, pp. 596–599.
[38] S. Haykin, *Neural Networks -A Comprehensive Foundation*. New York: Macmillan, 1994.

Yuan-Fu Liao received the B.S., M.S., and Ph.D. degrees from National Chiao Tung University (NCTU), Hsinchu, Taiwan, R.O.C., in 1991, 1993, and 1998, respectively.

He was a Postdoctoral Researcher with the Department of Communication Engineering, NCTU, from January 1999 to June 1999. He has been a Research Engineer with Philips Research East Asia, Taiwan, since September 1999. His major research interests are Mandarin speech recognition and the application of neural networks in speech processing.

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), Hsinchu, 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 with Telecommunication Laboratories, Chung-Li, Taiwan. He became an Associate Professor and then Professor with the Department of Communication Engineering, NCTU, in 1983 and 1990, respectively. He was Department Chairman from August 1985 to July 1988 and October 1991 to July 1993. His major research area is speech processing, especially in Mandarin speech recognition and text-to-speech.