Modular Recurrent Neural Networks for Mandarin Syllable Recognition

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

***Abstract*—A new modular recurrent neural network (MRNN)-based speech-recognition method that can recognize the entire vocabulary of 1280 highly confusable Mandarin syllables is proposed in this paper. The basic idea is to first split the complicated task, in both feature and temporal domains, into several much simpler subtasks involving subsyllable and tone discrimination, and then to use two weighting RNN's to generate several dynamic weighting functions to integrate the subsolutions into a complete solution. The novelty of the proposed method lies mainly in the use of appropriate *a priori* linguistic knowledge of simple *initial-final* structures of Mandarin syllables in the architecture design of the MRNN. The resulting MRNN is therefore effective and efficient in discriminating among highly confusable Mandarin syllables. Thus both the time-alignment and scaling problems of the ANN-based approach for large-vocabulary speech-recognition can be addressed. Experimental results show that the proposed method and its extensions, the reverse-time MRNN (Rev-MRNN) and bidirection MRNN (Bi-MRNN), all outperform an advanced HMM method trained with the MCE/GPD algorithm in both recognition-rate and system complexity.**

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

I. Introduction

In the past decade, many methods have been proposed for large-vocabulary speech-recognition. Among them, the most widely used and successful is the hidden Markov model (HMM)-based approach. In this approach, each class is represented by an HMM model trained usually by the maximum likelihood (ML) algorithm [1], which maximizes the within-class likelihood for each class without considering competition among hostile classes. The recent research trend in this approach is thus to eliminate this drawback by increasing the HMMs' discrimination abilities through the use of competitive training. Many discriminative training algorithms have been proposed. They include the maximum mutual information (MMI) algorithm [2], the maximum *a posteriori* (MAP) algorithm [3], and the minimum classification error/generalized probabilistic descent (MCE/GPD) algorithm [4], [5]. Simulation results have confirmed that better recognition performances can usually be achieved by these advanced HMM methods.

Along with the HMM approach, the artificial neural network (ANN)-based approach is also attractive because ANN's have the distinction of possessing high discrimination ability obtained through competitive training [6], [7]. Although many ANN-based methods have been proposed previously for speech recognition, only few of them are suitable for large-vocabulary applications because of the scaling and time-alignment problems. The scaling problem holds that an ANN structure that is effective for small-vocabulary speech-recognition is not guaranteed to also be effective for large-vocabulary applications simply because it has been scaled-up. The time-alignment problem results mainly from the use of ANN's, which are designed for static pattern recognition, to discriminate among dynamic speech patterns. Among all successful ANN-based methods for large-vocabulary speech recognition, hybrid HMM/ANN methods [8]–[10] and modular neural network (MNN)-based methods [11]–[12] are the most promising. We briefly review these two methods as follows.

Hybrid HMM/ANN methods incorporate ANN-emission probability estimators into the probabilistic frameworks of HMM's to relax some improper assumptions made by the HMM methods. They can partially overcome the weakness caused by noncompetitive training in conventional ML-trained HMM's. Many successful HMM/ANN methods [8]–[10] with performances better than those of the conventional HMM methods have been reported. Although they are all effective, there is some room for performance improvement. These methods only use ANN's as nonparametric probability function approximators, and do not use *a priori* knowledge of specific applications in their architecture designs to further improve recognition performances. Specifically, ANN's are only used for the frame-level discrimination, not phone segment-level or word-level discrimination, none of which uses full ability of ANN's. And ANN outputs cannot be directly used in HMM's as likelihood functions, they must be scaled according to *a priori* probabilities in order to fit HMM requirements [8]–[10]. Moreover, the scaling operation has no appropriate interpretations to minimize error rates in realizing the final goal of speech recognition. This may reduce its discrimination ability.

MNN-based methods use the "*divide-and-conquer*" principle to first decompose complicated large-vocabulary speech-recognition tasks into subtasks involving subset recognition or subunit recognition to then properly combine the results into a complete solution. One merit of MNN-based methods is that they provide interpretable and tractable ways to analyze the internal operations of the neural networks rather than simply taking them as black boxes. A key issue with MNN-based

Manuscript received June 20, 1997; revised September 4, 1998. This work was supported by National Science Council, R.O.C., under Contract NSC87-2213-E009-0027.

The authors are with the Department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, R.O.C.

Publisher Item Identifier S 1045-9227(98)08573-7.

methods is proper use of *a priori* domain knowledge in the design of the MNN architecture [13] in order to efficiently and effectively accomplish the main task of large-vocabulary speech-recognition. Currently, this key issue is still not well addressed in most MNN-based methods. Only few successful systems have been reported in the literature. They include the hierarchical mixture of experts (HME) [14]–[17], Meta-Pi networks [18], [19], one-word-one-net neural networks (OWON) [20], [21], and the hierarchical neural network (HNN) [22]. One main drawback of those methods is that they cannot solve the time-alignment problem by themselves. Besides, HME, Meta-Pi, and OWON only decompose the task in the feature domain, and ignore the rich information contained in the temporal domain. And OWON and HNN are inefficient, using too many subnetworks or parameters in their MNN's.

In this paper, a new speech recognition method based on a modular recurrent neural network (MRNN) is proposed for recognizing the entire vocabulary of 1280 highly confusable isolated Mandarin syllables. It decomposes this complicated task in both feature and temporal domains by first dividing it into several subtasks involving subsyllable and tone discriminations. And then uses two weighting RNN's to generate several dynamic weighting functions to integrate them. The novelty of the work is its proper use of appropriate *a priori* linguistic knowledge of Mandarin syllables' structure in the architecture design of the MRNN, which makes it effective and efficient in discriminating among highly confusable Mandarin syllables. The MRNN also solves time-alignment problem effectively by allowing temporal variations in speech signals and avoiding the time-consuming dynamic programming (DP) procedure. Moreover, all RNN' outputs are directly combined to form discriminant functions for all 1280 syllables. This makes it easy to design a discriminative training algorithm to optimize the MRNN with a goal of maximizing the syllable accuracy rate. The resulting MRNN is therefore suitable for Mandarin syllable recognition.

The organization of this paper is as follows. Section II provides background information about the task of recognizing the entire vocabulary of 1280 isolated Mandarin syllables. Section III presents the proposed method and its two extensions, the reverse-time MRNN (Rev-MRNN) and bidirectional MRNN (Bi-MRNN). A two-phase embedded training method with both subsyllable-level and syllable-level MCE/GPD algorithms is also described. Performance of the proposed method is evaluated in Section IV. Some conclusions are given in the last section.

II. Background Information and Problem Statement

Mandarin Chinese is a tonal language. There exist more than 80 000 words, each composed of from one to several characters. There are more than 10 000 commonly used characters, each pronounced as a monosyllable with a tone embedded in its fundamental frequency (F0) contour [23]. There are only about 1280 phonologically allowed syllables, and these comprise the set of all legal combinations of 411 base-syllables and five tones. Recognition of these 1280 syllables can thus be accomplished by combining the two subtasks of recognizing the 411 base-syllables and five tones. Although, from a vocabulary-size standpoint, the task seems to be relatively simple when compared with that of recognizing all English words, it is actually very difficult because the vocabulary contains many highly confusable syllables. This results mainly from the simple *initial-final* [or more roughly the consonant-vowel (C-V)] structure of base-syllables. Table I shows the phonetic structure of Mandarin syllables. The *initial* of a base-syllable, if it exists, consists only of a single consonant. There are a total of 22 *initials* including a null one (see Table II). The *final* is composed of three components including an optional preceding medial, a vowel nucleus, and an optional nasal ending. A total of 39 *final*s (see Table III) form all legal concatenations of three medials, 17 vowels, and two endings. With such a simple phonetic structure, all 411 base-syllables can be categorized into 39 confusing sets [23]. Like the English E-set, all base-syllables in each confusing set differ only in their *initial* consonants and are therefore difficult to distinguish among [24]–[25]. The A-set: [j-a], [ch-a], [sh-a], [tz-a], [ts-a], [s-a], [g-a], [k-a], [h-a], [d-a], [t-a], [n-a], [l-a], [b-a], [p-a], [m-a], [f-a] and the AN-set: [j-an], [ch-an], [sh-an], [r-an], [tz-an], [ts-an], [s-an], [g-an], [k-an], [h-an], [n-an], [b-an], [p-an], [m-an], [f-an] are two typical examples. Moreover, considerable cross-confusion among the base-syllables in these 39 confusing sets also exists [23]. For examples, [l-i-an], [l-u-an] and [l-iu-an] differ only in their medials; [k-en] and [k-eng] can only be distinguished by their ending nasals; and so on.

TABLE I
The Phonetic Structure of Mandarin Syllables

Tone			
(Initial)	*Final*		
(Consonant)	(Medial)	Vowel	(Ending)

TABLE II
22 Context-Independent *initials* and Their Corresponding Subgroups. Here ϕ_I Denotes a Null *initial*

initial								
		stop		affricate		fricative		
liquid	nasal	voiced	unvoiced	voiced	unvoiced	voiced	unvoiceed	null
								ϕ_I
	m	b	p					
							f	
				tz	ts		s	
l	n	d	t					
				j	ch	r	sh	
				ji	chi		shi	
		g	k				h	

TABLE III
39 *finals* AND THEIR CORRESPONDING SUBGROUPS. HERE ϕ_F DENOTES A NULL *final*

sub-group	*final*
ϕ_F-	ϕ_F
a-	a, ai, au, an, ang
o-	o, ou
e-	e, eh, ei, en, eng, er
i-	i, i-a, i-eh, i-ai, i-au, i-ou, i-an, i-en, i-ang, i-eng, i-er
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

Currently, the dominant technology for approaching the task is still HMM-based. Most HMM-based methods use the conventional within-class training algorithm to separately generate HMM models for each individual class without considering competition with hostile classes. Lee *et al.* proposed an isolated syllable-based dictation machine [23], [26]. The system was then upgraded into a continuous-speech device [27], [28]. Some HMM-based Mandarin speech-recognition methods that use the MCE/GPD discriminative training algorithm have recently been reported [29], [30].

Many ANN-based isolated Mandarin syllable-recognition methods have also been proposed. Wang *et al.* used a three-stage HNN to recognize all 1300 isolated Mandarin syllables [22]. Jou *et al.* used a OWON to recognize 408 isolated base-syllables [20]. They all used high-dimensional input feature vectors to represent whole-syllable (or subsyllable) patterns for recognition. Preprocessing to time-normalize input utterances was therefore needed. Chen *et al.* used a hierarchical recurrent neural network (HRNN) to recognize 54 highly confusable Mandarin syllables, all with nasal endings [31]. The HRNN recognizer is a preliminary version of the MRNN recognizer proposed in this study. It has a simpler structure and was used only in a pilot study of 54-base-syllable recognition. A hybrid approach that combines an ANN pattern recognizer and a DP search was also presented recently [32].

Many methods for tone recognition have been proposed. They include MLP-based methods for four-tone-recognition of isolated Mandarin syllables [33], and HMM-based methods for five-tone-recognition of continuous Mandarin speech [34]. A prosodic-model-based method [35] and a hybrid method [36] have also been proposed.

III. THE PROPOSED METHOD

An MRNN recognizer is proposed to discriminate among the entire set of 1280 isolated Mandarin syllables. Fig. 1 shows a block diagram of the MRNN recognizer. It decomposes the task of recognizing 1280 syllables into four subtasks including three discrimination subtasks and one integration subtask. The three discrimination subtasks do the lower-level within-group classifications for the three subsyllable groups of five tones, 100 *final*-dependent (FD) *initials*, and 39 *finals*. Here the 100 FD *initial*s form a complete, legal set expanded from these 22 context-independent *initial*s (see Table II) according to the seven broad-classes (see Table III) of their succeeding *final*s. They are used to partially compensate the intrasyllable coarticulation effect between the *initial* and *final* segments [27], [28], [30], [37]. The integration subtask does the upper-level among-group classifications, as well as the discriminant-function combinations required for final decision-making. As shown in Fig. 1, each discrimination subtask is accomplished using an RNN to generate within-group discriminant functions for all subsyllables belonging to the group it is associated with. The integration subtask is accomplished using two weighting RNN's and one discriminant-function combination module. The first weighting RNN generates three primary dynamic weighting functions for each of the three respective subsyllable groups of five tones, 100 FD *initials*, and 39 *finals*. The second weighting RNN generates nine additional secondary dynamic weighting functions for each of the nine *initial* subgroups that partition the set of 22 context-independent *initial*s (and the expansion set of 100 FD *initial*s) according to the manner of articulation. Table II shows that these nine *initial* subgroups include liquid, nasal, unvoiced and voiced stops, unvoiced and voiced affricates, unvoiced and voiced fricatives, and null. All these dynamic weighting functions are used in the discriminant-function combination module to combine subsyllable discriminant functions generated by those three discrimination RNN's in order to generate discriminant functions for the entire set of 1280 syllables. All five RNN's have similar three-layer simple recurrent structures with all

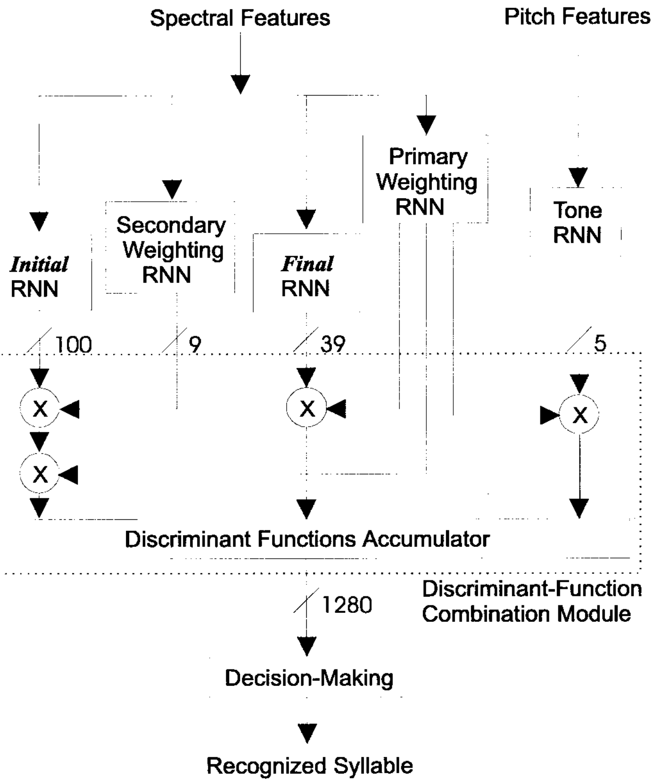


Fig. 1. A block diagram of the proposed MRNN syllable-recognizer.

outputs from their hidden layers being fed back to themselves as additional inputs [38]. All output-layer nodes in each RNN use linear output functions, rather than the more commonly-used sigmoid output functions. The operation of the MRNN recognizer is described in more detail below.

To recognize an input test utterance, a discriminant function is first defined for each of the 1280 syllables. For the pth syllable, which is composed of the ith tone, the jth FD *initial* (belonging to the lth *initial* subgroup), and the kth *final*, its discriminant function [5] can be expressed as

$$\begin{aligned} g_p(X_0^{L-1}) = \frac{1}{L}\sum_{n=0}^{L-1}[&O_i^T(X_n)O_T^W(X_n) \\ &+ O_j^I(X_n)O_l^{W_g}(X_n)O_I^W(X_n) \\ &+ O_k^F(X_n)O_F^W(X_n)], \\ &p = 0, \cdots, 1279 \end{aligned} \tag{1}$$

where $X_0^{L-1} = X_0, X_1, \cdots, X_{L-1}$ is the feature-vector sequence of a test utterance of length L; $O_i^T(X_n)$, $O_j^I(X_n)$, and $O_k^F(X_n)$ are, respectively, the ith output of the tone RNN, the jth output of the *initial* RNN, and the kth output of the *final* RNN; $O_T^W(X_n)$, $O_I^W(X_n)$, and $O_F^W(X_n)$ are the corresponding three primary dynamic weighting functions produced by the first weighting RNN; and $O_l^{W_g}(X_n)$ is the lth secondary dynamic weighting function produced by the second weighting RNN. The final decision rule then chooses the best candidate syllable according to the maximum discriminant function, i.e.,

$$\hat{p} = \arg \max_{p=0\sim 1279} g_p(X_0^{L-1}). \tag{2}$$

It is worth noting that the proposed method can be considered as a generalization of the conventional hybrid NN/DP method, which uses the set of much simpler weighting functions given below

$$O_T^W(X_n) = A \tag{3}$$

$$O_l^{W_g}(X_n)O_I^W(X_n) = \begin{cases} 1, & 0 \le n \le B \\ 0, & B < n \le L-1 \end{cases} \tag{4}$$

$$O_F^W(X_n) = \begin{cases} 0 & 0 \le n \le B \\ 1 & B < n \le L-1 \end{cases} \tag{5}$$

where A is a constant weight determined by try-and-error for properly combining the base-syllable and tone scores, and B is the *initial*/*final* boundary determined by DP search. We also note that the proposed MRNN has an architecture similar to those of the recurrent HME's proposed in [39]–[41], but they function in very different ways.

To train the MRNN recognizer, a special two-phase embedded training method with both subsyllable-level and syllable-level MCE/GPD algorithms [5] was incorporated in it. It uses the distributively-train-then-combine approach to first train all constituent RNN's separately in the first phase, and then combine all RNN's to form the whole MRNN syllable recognition system and fine-tune in the second phase. We chose this training method instead of direct whole-MRNN training because, in the first phase, all MRNN modules can be trained in parallel on only a small subset of the training data, thus speeding up the training process. Beside, the MRNN is a third-order system for which direct whole-MRNN training is potentially unstable, and may lead to the training process fail to converge. Moreover, direct whole-MRNN training would suffer from so-called spatial crosstalk (or interference) [11], and thus risk becoming trapped in local optima, resulting in slow convergence, poor generalization, or even learning failure. Below, the two phases of the proposed training algorithm are described in more detail.

In the first phase, distributive training, all three discrimination RNN's and the two weighting RNN's are trained separately. Preprocessing is first used to segment all training syllable utterances into *initial* and *final* segments. Segmentation is realized by using the HMM-based segmentation method to find the *initial*/*final* boundary for each training utterance. Then, the boundary is relaxed allowing the *initial* and *final* segments to overlap by several frames, after which the subsyllable-level MCE/GPD algorithms [5], [31] are used to train the *initial*, *final*, and tone discrimination RNN's separately on the *initial* segments, the *final* segments, and the voiced segments, respectively. The voiced segment of an utterance is defined as that portion of the pitch period that can be reliably detected. At the same time, the two weighting RNN's are being independently trained by the error backpropagation (EBP) algorithms [42] to generate appropriate dynamic weighting functions for the three subsyllable groups and nine *initial* subgroups. "$0-1$" output target functions determined according to the segmentation results are used for all 12 dynamic weighting functions. It is worth noting that the coarticulation effect between the *initial* and *final* segments on syllable recognition can be partially compensated for by using overlapped *initial* and *final* segments to train the *initial* and *final* RNN's during the first training phase.

After training the five RNN's, we then execute the second phase, combining training. All five RNN's are combined using the discriminant-function combination module and then fine-tuned using a "bootstrapping" procedure embedded within a syllable-level MCE/GPD algorithm according to the discriminant function defined in (1). In the bootstrap fine-tuning procedure, all constituent RNN's of the MRNN was distributed into three parts and retrained part-by-part, i.e., we select one part at a time for retraining and freeze the weights of the other parts. Detailed derivations of the syllable-level MCE/GPD training algorithms are given in the Appendix. The reason for using a bootstrapping procedure rather than direct whole-MRNN retraining is also to avoid encountering instability in the training process. We note that the "$0-1$" bounds of all dynamic weighting functions, set as learning targets during first-phase training, are now relaxed and freed of any manual control during the second-phase training. The ultimate level that a dynamic weighting function can reach is automatically determined by the syllable-level MCE/GPD algorithm. This increases the discrimination capacity of the MRNN syllable-recognizer by placing special emphases on the most distinguishing parts of input test utterances for each candidate syllable [24]–[25].

The basic MRNN syllable recognizer described above is improved by first adding an additional Rev-MRNN syllable recognizer, and then combining these two MRNN's into a bidirectional syllable recognizer, i.e., the Bi-MRNN. The Rev-MRNN is the same as the basic MRNN except that the time scale was reversed, and is designed to utilize the right-context information of speech signal instead of the left-context information in the basic MRNN. Therefore, the Bi-MRNN can explore the context-information of a speech signal in both directions to further improve its performance [43]. Many combination methods [43] can be used to form the Bi-MRNN syllable recognizer. In this study, we used a simple combination method that directly averages the syllable discriminant functions of the two MRNN syllable recognizers to form the final syllable discriminant functions. The final syllable discriminant function for the pth syllable can thus be expressed by

$$\begin{aligned} g_p^{\text{Bi-MRNN}}(X_0^{L-1}) \\ &= \tfrac{1}{2} \cdot \{g_p^{\text{MRNN}}(X_0^{L-1}) + g_p^{\text{Rev-MRNN}}(X_0^{L-1})\} \\ &\quad p = 0, \cdots, 1279. \end{aligned} \tag{6}$$

The Bi-MRNN are also trained by the above mentioned two-phase embedded training method according to the discriminant function defined in (6). Detailed derivations of the training algorithms are similar to those given in the Appendix. Note that although a simple averaging combination method is used here, the basic MRNN and Rev-MRNN recognizers are of unequal importance in influencing the final syllable-recognition decision. The syllable-level MCE/GPD algorithm will fine-tune the Bi-MRNN syllable recognizer to automatically adjust the dynamic weighting functions of the two MRNN's to their appropriate values.

Several distinctive characteristics of the MRNN recognizer are given below.

- It is a frame-based recognition system. Recognition feature vectors are fed in frame-by-frame. No preprocessing to time-normalize input test utterances is needed. The input layers of all RNN's in the MRNN are therefore much smaller as compared with those using whole-utterance input patterns [20]–[22].
- The short-term context-dependencies of input feature vectors are explicitly invoked during recognition testing by using RNN's to discriminate among subsyllables. As mentioned above, all RNN's have similar three-layer structures with all outputs from hidden layers being fed back to themselves as additional inputs. Thus each RNN functions as a dynamic system with the outputs of its hidden layer depending at any given time on a complex aggregate of all previous inputs [9], [38].
- The architecture of the MRNN recognizer has been designed to conform to Mandarin Chinese's syllabic phonetic structure. Subsyllables with similar acoustic properties are grouped together and recognized using the same RNN. This makes each individual RNN an expert classifier of associated subsyllable subsets through concentration on the distinguishing parts of constituent subsyllables.
- *Initials*, which are the most highly confusable subsyllables, receive special care in the MRNN recognizer: 100 FD *initials* are used as recognition units, and then several sets of dynamic weighting functions are applied to increase their distinguishablility.
- The MRNN is architecturally homogeneous. Except for the discriminant-function combination module, it is all made up of RNN's.
- The time-alignment considerations have been resolved via the use of the weighting RNN's. Thus, no DP's are needed to calculate syllable discriminant scores [see (1) and (6)]. This greatly simplifies the recognition process.

As compared with previously proposed methods, the proposed MRNN recognizer has the following distinct advantages.

- It is more effective on Mandarin syllable-recognition than the HMM and hybrid HMM/ANN methods because it considers not only the frame-level but also subsyllable- and syllable-level competition among hostile syllables during the recognition process. Besides, it lets each frame of the input test speech signal contribute uniquely to the final recognition decision via the use of several sets of dynamic weighting functions.
- It is also more sophisticated in that it uses the information in the transient area between the *initial* and *final* segments for both *initials* and *finals* discriminations.
- The time-alignment requirement inherent in applying ANN's to the task of large-vocabulary speech recognition is solved. Temporal variations in test speech patterns are absorbed partially by the discrimination RNN's and partially by the weighting RNN's.

IV. Evaluations

The performance of the proposed method was evaluated using a multispeaker speech-recognition task. A database containing utterances of all 1280 isolated Mandarin syllables was used in the following experiments. Each syllable was pronounced ten times by eight male and two female speakers with a Taiwan accent. Seven repetitions were used for training and the remaining three for testing. All speeches were directly digitally recorded in a laboratory using a personal computer with a 16-bit Sound Blaster card and a head-set microphone. A sampling rate of 16 kHz was used. All speech signals were then preprocessed to extract features for recognition. During preprocessing, signals were first preemphasized with a high-pass filter $1 - 0.95z^{-1}$, then, a 14th-order LPC analysis was performed on every 20-ms Hamming-windowed frame with 10-ms frame shifts. Fourteen liftered LPC-derived cepstral coefficients and their first-order time derivatives were then computed. The first-order and second-order time derivatives of the log energy, as well as the zero-crossing rate, were also computed for each frame. All 31 parameters were then taken as input features for base-syllable recognition. Five parameters were used for tone recognition: log energy, delta log energy, peak normalized auto-correlation function, pitch period, and delta pitch period. The pitch period was automatically detected using the simple inverse filter tracking (SIFT) algorithm [44] without any modification. The window size for pitch analysis was 40 ms with 10-ms shifts.

TABLE IV
RECOGNITION RESULTS OF THE DISTRIBUTIVE-TRAINING PROCEDURE FOR THE NINE *initial* SUB-GROUPS, 100 FD *initials*, 39 *finals*, AND FIVE TONES

	# of hidden neurons	9 *initial* sub-groups	100 FD *initials*	39 *finals*	5 tones
RNN	30	85.0%			88.3%
	60		76.9%	88.6%	
	120		80.6%	89.5%	
	150		80.9%	90.0%	
Rev-RNN	30	83.4%			87.5%
	60		77.2%	88.9%	
	120		80.3%	90.3%	
	150		82.0%	90.6%	

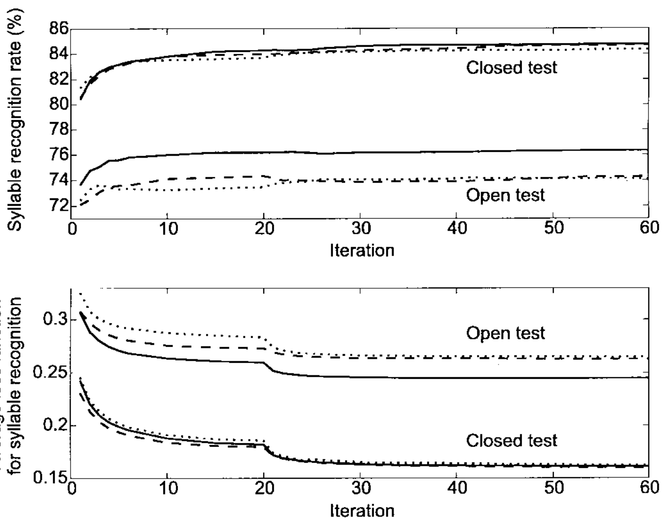


Fig. 2. The learning curves of the three schemes of the proposed MRNN recognizer for 1280 Mandarin syllables (solid line: Bi-MRNN, dashed line: MRNN, dotted line: Rev-MRNN).

A. Experimental Results

We started by examining the performance of the proposed MRNN syllable-recognizer. The effect of first-phase distributive training was examined first. All input-syllable utterances were presegmented into *initial* and *final* segments with 100-ms overlaps. All five RNN's were then trained separately with target information being set according to the given segmentations of all training utterances. The numbers of neurons used in the hidden layers of the primary weighting RNN, the secondary weighting RNN, and the tone discrimination RNN were all set to fixed values of 30. By contrast, several values for the numbers of hidden neurons were tested for the *initial* and *final* discrimination RNN's, including 60, 120, and 150 hidden neurons. All five RNN's were trained in parallel on several workstations for 100 iterations. The recognition results for the *initial-*, *final-*, and tone-discrimination RNN's are shown in Table IV. The best recognition rates achieved for 100 FD *initials,* 39 *finals,* and five tones were 80.9, 90.0, and 88.3%, respectively. We also report the performance of the secondary weighting RNN on the nine *initial* subgroup classification in Table IV. A fair classification rate of 85.0% was obtained. We note that the classification performance of the secondary weighting RNN is given for reference only and is not important to the MRNN-based syllable recognition. This is because the secondary weighting RNN is only trained by the EBP algorithm using frame-based "$0 - 1$" target functions with a goal to provide proper dynamic weighting functions for the *initial* part of the testing utterance. Besides, it will be fine-tuned in the second-phase training.

We then combined the five RNN's to form the basic MRNN 1280-Mandarin-syllable-recognizer and retrained it using the bootstrapping procedure. As mentioned above, all constituent RNN's of the MRNN were distributed into three parts and fine-tuned part-by-part. The fine-tuning sequence started with the secondary weighting RNN, continued with the primary weighting RNN, and ended with the *initial-*, *final-*, and tone-discrimination RNN's. Twenty iterations were performed in each of the three retraining steps. Fig. 2(a) and (b) show, respectively, the syllable recognition-rate learning curves and the average loss functions for syllable recognition [defined in (A.2)] achieved in both closed and open tests. Fig. 2(a) shows that the recognition rates for closed and open tests both increased very rapidly at the beginning of the first retraining step, and then went into saturation. Conversely, the two average loss functions decreased very rapidly during the first several iterations of the first retraining step, and then became saturated. They then decreased again at the beginning of the second retraining step, and again went into saturation quickly. These results show that the retraining gain in recognition performance decreased as we proceeded with the bootstrapping procedure. A recognition rate of 74.2% was finally obtained. It is worth noting that the sum of the recognition rate and the average loss function equaled approximately 1.0 when the training process converged. This outcome is the same as that obtained in [24] and [25]. Thus the effectiveness of the bootstrapped retraining procedure has been confirmed.

The performances of the Rev-MRNN and Bi-MRNN recognizers on the 1280 Mandarin syllables were then examined. The same two-phase training algorithm was used to train these two systems. The training procedure for the Rev-MRNN syllable recognizer was the same as that used to train the basic MRNN syllable recognizer, except that the time scale was reversed. Five reverse-time RNN's were trained separately during the first phase (see Table IV), then combined and fine-tuned during the second phase. The training procedure was similar for the Bi-MRNN syllable recognizer. We first trained the ten RNN's separately, five in the basic MRNN and five in the Rev-MRNN, in the first-phase training (see Table IV), then combined them and used the same three-step bootstrap retraining procedure to fine-tune them during second-phase training. Note that in each step of the bootstrapping procedure, the same parts of the two MRNN's were retrained simultaneously. The learning curves of these two MRNN syllable recognizers

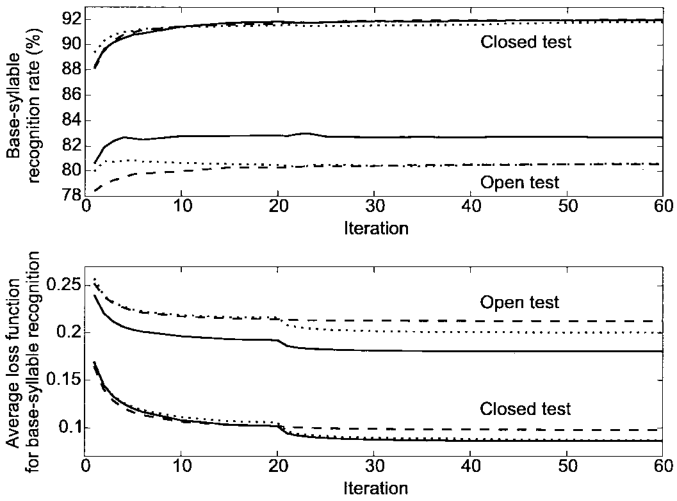


Fig. 3. The learning curves of the three schemes of the proposed MRNN recognizer for 411 Mandarin base-syllables (solid line: Bi-MRNN, dashed line: MRNN, dotted line: Rev-MRNN).

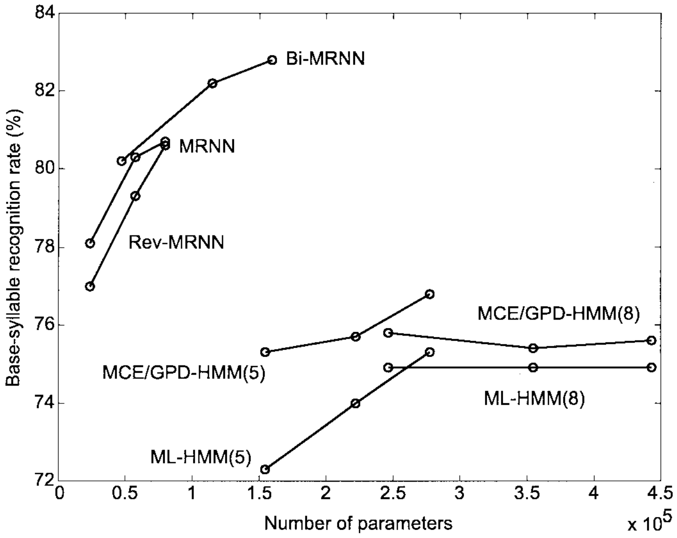


Fig. 4. Recognition results of the ML-, MCE/GPD-trained HMM and the proposed MRNN methods for 411 Mandarin base-syllables. Here HMM(x) means the maximum mixture number in each state is set to x.

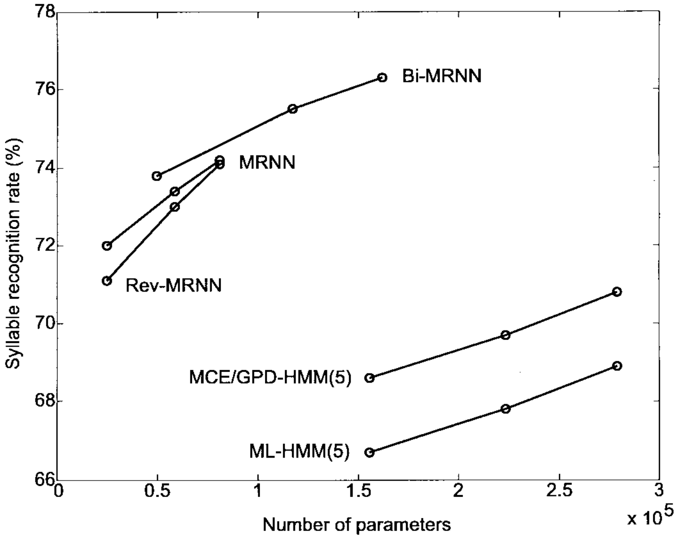


Fig. 5. Recognition results of the ML-, MCE/GPD-trained HMM and the proposed MRNN methods for 1280 Mandarin syllables. Here HMM(x) means the maximum mixture number in each state is set to x.

are also shown in Fig. 2. As the figure shows, the learning processes of these two syllable recognizers were very similar to that of the basic MRNN system. Recognition rates of 74.1 and 76.3% were achieved by the Rev-MRNN and Bi-MRNN syllable recognizers, respectively. The performance of the Rev-MRNN syllable recognizer was comparable to that of the basic system, but the performance of the bidirectional system was better. This shows that the complementary characteristics of the Rev-MRNN to the basic MRNN on Mandarin syllable recognition [43].

Due to its relative importance, recognition of 411 Mandarin base-syllables was also examined using the MRNN, Rev-MRNN, and Bi-MRNN. Fig. 3 shows the learning curves for these three recognizers. Similar learning processes can be observed in the figure. Respective recognition rates of 80.7, 80.6, and 82.8% were achieved by the three systems.

For a performance comparison, the continuous-density HMM method using the same subsyllable recognition units to recognize the 411 Mandarin base-syllables was also tested. Three combinations of state numbers for *initial*s and *final*s were used, including (3,5), (4,8), and (5,12). Here (x, y) denotes that the state numbers of *initial* and *final* HMM models are set to x and y, respectively. All HMM models had left-to-right structures and no states were skipped. The observation features in each state were modeled by a partitioned-mixture Gaussian distribution. The mixture number in each state varied depending on the number of training data, but a maximum mixture number of five or eight was set. The best recognition rate was achieved when there were (5,12) states and five mixtures. It was only 75.3% (see Fig. 4). For a fairer comparison, these ML-trained HMM's were then further refined using the syllable-level MCE/GPD training algorithm [5]. The best recognition rate increased to 76.8% (see Fig. 4), but this was still far below that achieved by the proposed Bi-MRNN base-syllable recognizer. To compare performance on recognition of the 1280 Mandarin syllables, we combined the base-syllable HMM's with the same RNN tone-recognizer used in the basic MRNN system to form a hybrid syllable-recognizer. This hybrid system was also further refined using the syllable-level MCE/GPD training algorithm. Recognition rates of 68.9 and 70.8% (see Fig. 5) were achieved by the ML-trained and MCE/GPD-trained HMM's, respectively. But they were still far lower than that obtained by the proposed Bi-MRNN system. Thus the proposed MRNN-based method outperformed the HMM method.

We then examined the computational efficiencies of these five methods. Rough comparisons based on total numbers of parameters used by these five methods were made. The results are also plotted in Figs. 4 and 5 for the two cases involving recognition of the 411 base-syllables and 1280 syllables, respectively. These two figures show that all three schemes in the proposed method used fewer parameters than

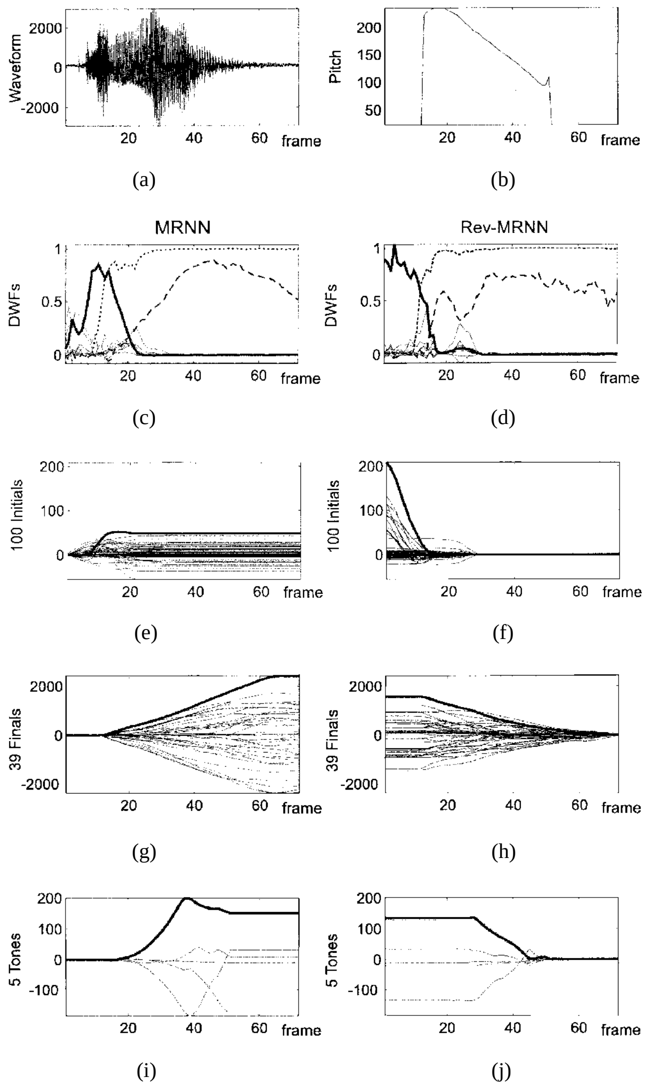


Fig. 6. Typical responses of the Bi-MRNN after the first-phase training: (a) the waveform of the syllable /ji-i-ang-4/ and (b) its pitch contour; the dynamic weighting functions (DWF's) for (c) the basic (forward) MRNN and (d) the Rev-MRNN (bold solid line: product of the two DWF's for the desired *initial* subgroup, solid lines: products of the two DWF's for other *initial* subgroups, dotted line: DWF for *final,* dash line: DWF for tone); the cumulative weighted discriminant functions of (e) the 100 FD *initials*, (g) 39 *finals*, and (i) five tones for the basic MRNN; and those of (f) the 100 FD *initials*, (h) 39 *finals*, and (j) five tones for the Rev-MRNN (bold solid lines: the desired *initial*, *final*, and tone classes).

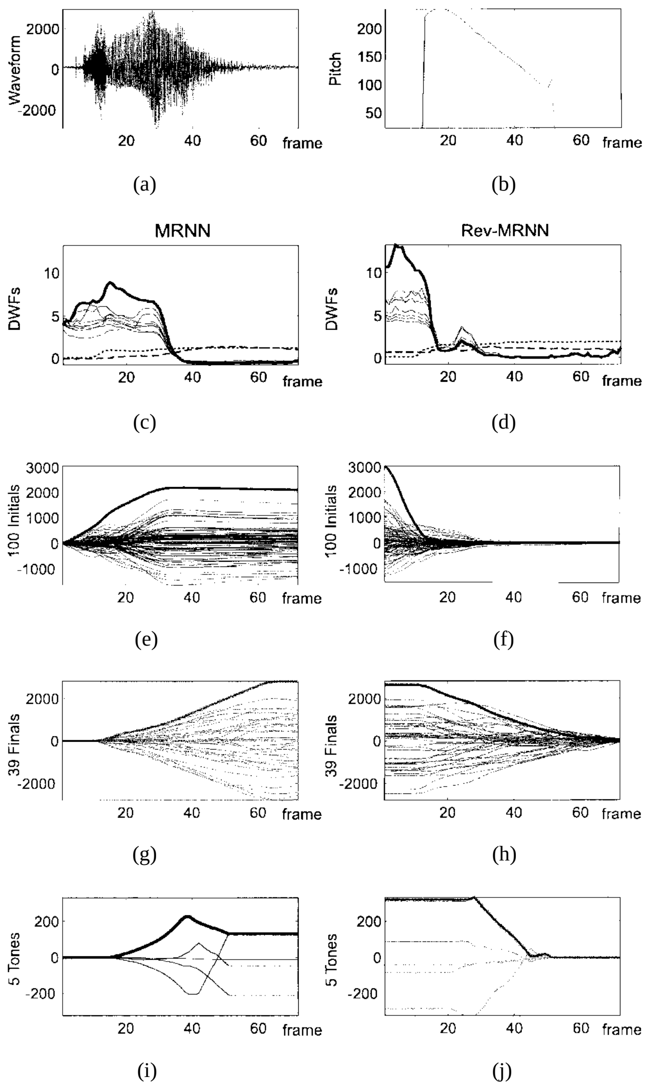


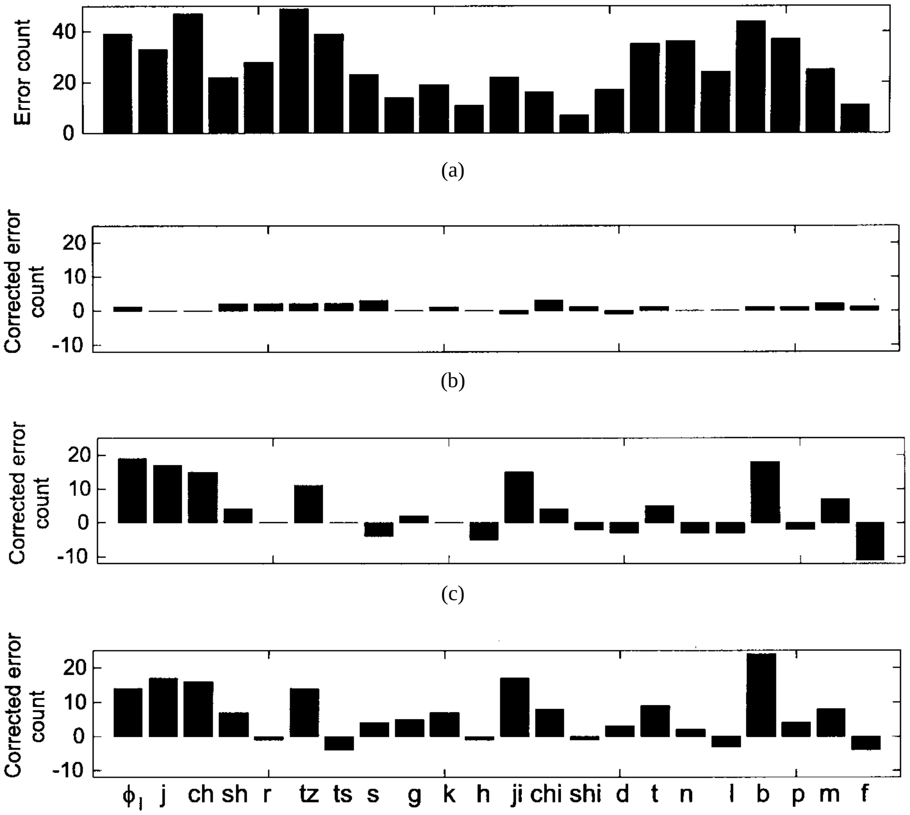
Fig. 7. Typical responses of the Bi-MRNN after the second-phase training: (a) the waveform of the syllable /ji-i-ang-4/ and (b) its pitch contour; the dynamic weighting functions (DWF's) for (c) the basic (forward) MRNN and (d) the Rev-MRNN (bold solid line: product of the two DWF's for the desired *initial* subgroup, solid lines: products of the two DWF's for other *initial* subgroups, dotted line: DWF for *final*, dash line: DWF for tone); the cumulative weighted discriminant functions of (e) the 100 FD *initials*, (g) 39 *finals*, and (i) 5 tones for the basic MRNN; and those of (f) the 100 FD *initials*, (h) 39 *finals*, and (j) 5 tones for the Rev-MRNN (bold solid lines: the desired *initial*, *final*, and tone classes).

the two HMM methods. Beside, no DP's were needed in the recognition tests of the three schemes in the proposed method. They were thus all much more efficient than the HMM method.

B. Analysis of the MRNN Operation

A detailed analysis of the operation of the proposed MRNN was then performed. This was worthwhile since it yielded a better understanding of its behavior. Typical responses of the Bi-MRNN, obtained before and after we applied the second-phase combining retraining procedure, are shown in Figs. 6 and 7, respectively. These two figures show several distinctive properties of the two-phase training process.

- As shown in Fig. 6(c) and (d), the active levels of the *initial*, *final*, and tone dynamic weighting functions are all the same after the first-phase distributive training. Here, the active level of an output response is generally defined as the approximate level reached when an input utterance of the class associated with it is received. On the contrary, as shown in Fig. 6(e)–(j), the active levels of the cumulative weighted discriminant functions for 100 FD *initials*, 39 *finals*, and five tones are quite different. The two active levels for 39 *finals* are much larger than all others. This means that, after the first-phase training, the recognition results are dominated primarily by the *final* part of the test utterance. This is reasonable because the *final* part of a test utterance contains, in average, much more input frames and the three weighting functions have approximately equal active levels.



(a)

(b)

(c)

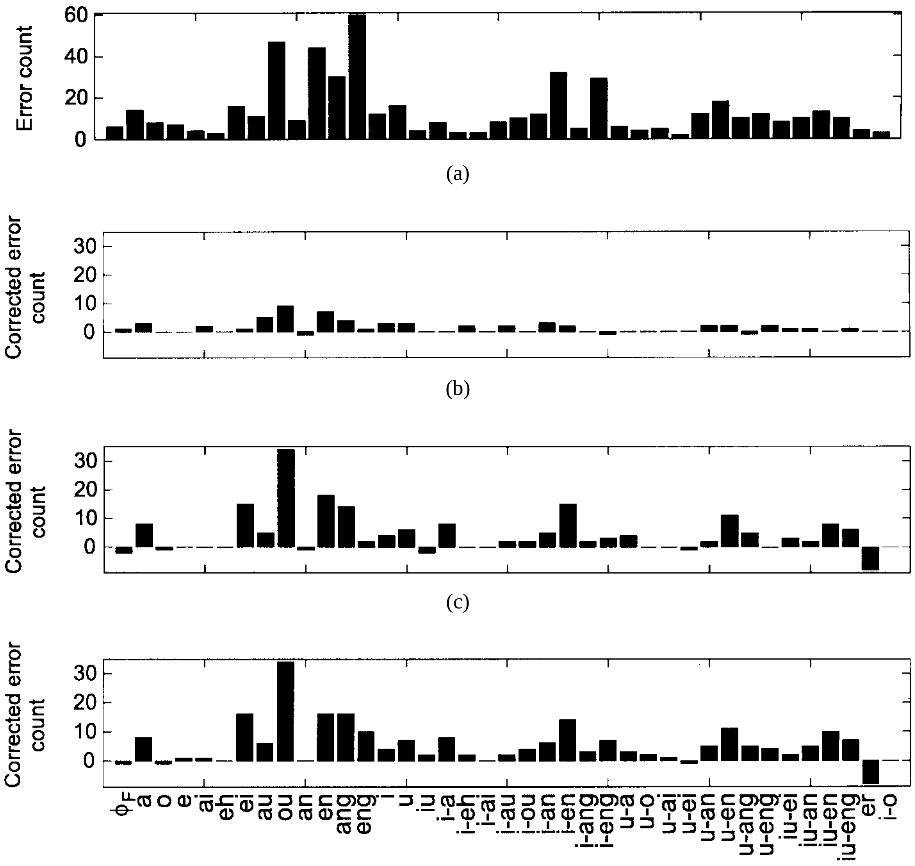
(d)

Fig. 8. Error analysis for 22 context-independent *initials*: (a) error counts for the ML-trained HMM; and corrected error counts compared with the ML-trained HMM for (b) MCE/GPD-trained HMM, (c) MRNN, and (d) Bi-MRNN.

- As shown in Fig. 7(c) and (d), the active levels of the *initial* dynamic weighting functions are much higher than those of the *final* and tone after the second-phase training. The contribution of the *initial* part of the test utterance to the final recognition decision is therefore emphasized. This result is consistent with those obtained in [24] and [25] for the recognition of English E-set. Actually, as shown in Fig. 7(e)–(h), the active levels of the cumulative weighted discriminant functions for 100 FD *initial*s and 39 *finals* are approximately the same. So they contribute almost equally to the recognition of 411 base-syllables regardless of the inherent difference in their average durations.
- Fig. 7(c) and (d) shows that the active levels of the *initial* and *final* dynamic weighting functions were automatically adjusted to their appropriate levels after the second-phase training. The *initial* dynamic weighting functions truly emphasize the *initial* parts of input utterances and suppresses the *final* parts, and the *final* dynamic weighting function do the opposite. So together, they provide proper information about *initial*/*final* segmentation, thus simplifying the recognition process by directly combining the weighted discriminant functions without invoking a DP procedure.
- It can be found from Fig. 7(e)–(j) that the cumulative weighted discriminant functions of the *initial*, *final*, and tone in the basic MRNN have different active levels from their counterparts in the Rev-MRNN after the second-phase training. For both the *initial* and tone, they are much higher in the Rev-MRNN than in the basic MRNN. For the *final*, it is slightly higher in the basic MRNN than in the Rev-MRNN. This shows that the basic MRNN and Rev-MRNN use different context information in their syllables' discrimination. So, they can complement to each other.

C. Error Analyzes and Discussions

Lastly, some error analyzes were made. Statistics on *initial* and *final* recognition errors made by the ML-trained HMM method, the MCE/GPD-trained HMM method, the basic MRNN recognizer, and the Bi-MRNN recognizer were calculated for comparison, and are shown in Figs. 8 and 9. We note that in order to show the results more clearly, the errors for the 100 FD *initials* are grouped together to form error sets of 22 context-independent *initials*. Some observations concerning these two figures can be made. First, Fig. 8(a) shows that the most serious *initial* errors occurred on the *initial*s in $\{\phi_I$, [j], [ch], [tz], [ts], [t], [n], [b], [p]$\}$. Among them, $\{$[j], [tz]$\}$, $\{$[ch], [ts]$\}$, and $\{$[b], [p]$\}$ were three most confusable pairs. The first two pairs are retroflex sound/nonretroflex sound pairs. And [b], [p], and [t] are stop sounds that are well known to be difficult to recognize. Second, as Fig. 9(a) shows, the most serious *final* errors occurred on the *final*s in $\{$[ou], [en], [ang], [eng], [i-en], [i-eng]$\}$. Among them, $\{$[en], [eng]$\}$ and $\{$[i-en], [i-eng]$\}$ were the two most confusable pairs



(a)

(b)

(c)

(d)

Fig. 9. Error analysis of 39 context-independent *finals*: (a) error counts for the ML-trained HMM; and corrected error counts compared with the ML-trained HMM for (b) MCE/GPD-trained HMM, (c) MRNN, and (d) Bi-MRNN.

which are ended with nasals. The difficult to distinguish those retroflex/nonretroflex pairs and nasal ending pairs is partially due to the accent of Taiwanese. Third, in comparing Fig. 8(c), (d) and Fig. 9(c), (d), with Figs. 8(b) and 9(b), we find that both the basic MRNN and Bi-MRNN recognizers were better to correct *initial* and *final* errors than the MCE/GPD-trained HMM method. In particular, almost half the errors occurring on the most confusable *initial*s in {[j], [tz], [ch], [b], ϕ_I} and *final*s in {[ou], [ang], [en], [i-en]} made by the ML-trained HMM method were corrected in both the MRNN and Bi-MRNN recognizers.

V. Conclusions

In this paper, a new MRNN-based method for recognizing all legal Mandarin syllables has been described. The scaling and time-alignment problems encountered when applying ANN-based pattern recognition approaches to large-vocabulary speech recognition have been solved in unique fashion. The novelty of the proposed method mainly lies in the use of *a priori* domain knowledge about Mandarin syllable phonetic structures in designing the MRNN architecture. So it is superior to many current methods, including the advanced MCE/GPD-trained HMM method and some ANN-based methods. A further study to extend it to continuous Mandarin speech-recognition is a worthwhile future project [46]–[48].

Appendix
The Syllable-Level MCE/GPD Algorithm for MRNN Training

The procedure for applying the syllable-level MCE/GPD algorithm to MRNN training is described below. Although, we only discuss the algorithm for the basic MRNN here. The deriving of the training algorithms for the Rev-MRNN and Bi-MRNN are almost the same. First, based on the syllable discriminant function defined in (1), a mis-classification measure [5] for a input utterance X_0^{L-1} of the pth syllable is defined as

$$d_p(X_0^{L-1}) = -g_p(X_0^{L-1}) + g_{p^*}(X_0^{L-1}) \tag{A.1}$$

where p^* is the most probable incorrect syllable. A loss function for syllable recognition $J(d_p)$ is then defined to evaluate the cost of the current decision for the syllable. The loss function should be a monotonically increasing, differentiable function. If a well-approximated "$0 - 1$" function is used for $J(d_p)$, the average loss function $\boldsymbol{J}$ of a N utterances set

$$\boldsymbol{J} = \frac{1}{N} \sum_{X_0^{L-1}} J(d_p(X_0^{L-1})) \tag{A.2}$$

will approximately represent the syllable recognition error rate. In this study, the sigmoidal function

$$J(d_p) = \frac{1}{1 + e^{-\nu d_p}} \tag{A.3}$$

was chosen as the loss function. Here ν is a scalar for controlling the rate of weight adjustment. It is clear that the above loss function will force the training to emphasize utterances located a short distance from the decision boundary, and the scalar ν serves to scale that distance. The objective of the MCE/GPD algorithm is to recursively adjust the weights of the MRNN to achieve minimization of $\boldsymbol{J}$. A three-step bootstrapping procedure is used to sequentially adjust the weights of the five constituent RNN's. Thus, for each input utterance X_0^{L-1} of the N utterances set, the corresponding amount of weight change in these RNN's can be expressed via the MCE/GPD algorithm as

- Step 1:

$$\begin{aligned}\triangle W^{W_g} =& -\eta_n \cdot J'(d_k(X_0^{L-1}))\\ &\cdot \frac{1}{L}\sum_{n=0}^{L-1}\left\{\left[-\frac{\partial O_l^{W_g}(X_n)}{\partial W^{W_g}}O_j^I(X_n)\right.\right.\\ &\left.\left.+\ \frac{\partial O_{l^*}^{W_g}(X_n)}{\partial W^{W_g}}O_{j^*}^I(X_n)\right]O_I^W\right\},\\ &\text{for the secondary weighting RNN} \qquad \text{(A.4)}\end{aligned}$$

- Step 2:

$$\begin{aligned}\triangle W^{W} =& -\eta_n \cdot J'(d_k(X_0^{L-1}))\cdot \frac{1}{L}\sum_{n=0}^{L-1}\\ &\cdot\left\{\frac{\partial O_T^W(X_n)}{\partial W^W}[-O_i^T(X_n)+O_{i^*}^T(X_n)]\right.\\ &+\frac{\partial O_I^W(X_n)}{\partial W^W}[-O_l^{W_g}(X_n)O_j^I(X_n)\\ &+O_{l^*}^{W_g}(X_n)O_{j^*}^I(X_n)]\\ &\left.+\ \frac{\partial O_F^W(X_n)}{\partial W^W}[-O_k^F(X_n)+O_{k^*}^F(X_n)]\right\}\\ &\text{for the primary weighting RNN} \qquad \text{(A.5)}\end{aligned}$$

- Step 3:

$$\begin{aligned}\triangle W^{T} =& -\eta_n \cdot J'(d_k(X_0^{L-1}))\\ &\cdot \frac{1}{L}\sum_{n=0}^{L-1}\frac{\partial[-O_i^T(X_n)+O_{i^*}^T(X_n)]}{\partial W^T}O_T^W(X_n),\\ &\text{for the tone RNN} \qquad \text{(A.6)}\end{aligned}$$

$$\begin{aligned}\triangle W^{I} =& -\eta_n \cdot J'(d_k(X_0^{L-1}))\cdot \frac{1}{L}\sum_{n=0}^{L-1}\\ &\cdot\frac{\partial[-O_l^{W_g}(X_n)O_j^I(X_n)+O_{l^*}^{W_g}(X_n)O_{j^*}^I(X_n)]}{\partial W^I}\\ &\cdot O_I^W(X_n),\\ &\text{for the } initial \text{ RNN} \qquad \text{(A.7)}\end{aligned}$$

$$\begin{aligned}\triangle W^{F} =& -\eta_n \cdot J'(d_k(X_0^{L-1}))\cdot \frac{1}{L}\sum_{n=0}^{L-1}\\ &\cdot\frac{\partial[-O_k^F(X_n)+O_{k^*}^F(X_n)]}{\partial W^F}O_F^W(X_n),\\ &\text{for the } final \text{ RNN} \qquad \text{(A.8)}\end{aligned}$$

where η_n is the learning rate at the nth iteration, i^*, j^*, and k^* are, respectively, the most probable incorrect tone, FD *initial* (belonging to the l^*th *initial* subgroup), and *final*. The five derivative terms in the equations above can actually be computed through applications of the chain rule as suggested in [42]. The scheme for choosing a proper learning rate can be found in [49].

It is worth noting that in (A.5) the accumulation of the term $\partial O_F^W(X_n)/\partial W^W$ is weighted by $[-O_k^F(X_n)+O_{k^*}^F(X_n)]$, which is the difference between the outputs of the *final* RNN for the correct *final* and the most probable incorrect *final* at time n. This enables the primary weighting RNN to be trained to emphasize the parts of the input utterances that are important in distinguishing between the correct *final* and the most probable incorrect one. The same consideration can also be applied to the weight adjustments in the primary weighting RNN's for *initial* and tone, and in the secondary weighting RNN for *initial* subgroup. And from (A.8), the accumulation of the term $\partial[-O_k^F(X_n)+O_{k^*}^F(X_n)]/\partial W^F$ is windowed by the dynamic weighting function $O_F^W(X_n)$. This makes the *final* RNN focus on discriminating among the *final* parts of input utterances. The same consideration can also be applied to the weight adjustments of the *initial* and tone RNN's.

ACKNOWLEDGMENT

The authors would like to thank the editor and reviewers for their help in improving the paper.

REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] Y. Normandin, "Maximum mutual information estimation of hidden Markov models," *Automatic Speech And Speaker Recognition—Advanced Topics*. Boston, MA: Kluwer, 1996, pp. 57–82.

[3] C. H. Lee and J. L. Gauvain, "Bayesian adaptive learning and MAP estimation of HMM," *Automatic Speech And Speaker Recognition—Advanced Topics*. Boston, MA: Kluwer, 1996, pp. 83–108.

[4] S. Katagiri, C. H. Lee, and B. H. Juang, "New discriminative training algorithms based on the generalized probabilistic descent method," in *Proc. IEEE Neural Networks for Signal Process. (NNSP)*, 1991, pp. 299–308.

[5] B. H. Juang, W. Chou and C. H. Lee, "Statistical and discriminative methods for speech recognition," *Automatic Speech And Speaker Recognition—Advanced Topics*. Boston, MA: Kluwer, 1996, pp. 109–132.

[6] L. Niles, H. Sliverman, G. Tajchman, and M. Bush, "How limited training data can allow a neural network to outperform an 'optimal' statistical classifier," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 1989, pp. 17–20.

[7] H. Bourland, and C. J. Wellekens, "Links between Markov models and multilayer perceptrons," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, no. 12, pp. 1167–1178, 1990.

[8] H. Bourlard and N. Morgan, *Connectionist speech recognition—A hybrid approach*. Boston, MA: Kluwer, 1994.

[9] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, pp. 298–305, Mar. 1994.

[10] D. Kershaw, T. Robinson, and S. Renals, "The 1995 ABBOT LVCSR system for multiple unknown microphones," in *Int. Conf. Spoken Language Processing (ICSLP)*, 1996.

[11] 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," *Cognitive Sci.*, vol. 15, pp. 219–250, 1991.

[12] B. L. M. Happel and Jacob M. J. Murre, "Design and evolution of modular neural-network architectures," *Neural Networks*, vol. 7, nos. 6/7, pp. 985–1004, 1994.

[13] R. E. Jenkins and B. P. Yuhas, "A simplified neural-network solution through problem decomposition: The case of the truck backer-upper," *IEEE Trans. Neural Networks*, vol. 4, pp. 718–720, July 1993.

[14] M. I. Jordan and R. A. Jacobs, “Hierarchical mixtures of experts and the EM algorithm,” *Neural Comput.*, vol. 6, pp. 181–214, 1994.

[15] Y. Zhao, R. Schwartz, J. Sroka, and J. Makhoul, “Hierarchical mixtures of experts methodology applied to continuous speech recognition, in *Advances in Neural Inform. Processing Syst. 7 (NIPS’7)*, 1995, pp. 859–865.

[16] D. Kershaw, T. Robinson, and M. Hochberg, “Context-dependent classes in a hybrid recurrent network-HMM speech recognition system,” in *Advances in Neural Inform. Processing Syst. 8 (NIPS’8)*, 1996, pp. 750–756.

[17] J. Fritsch, M. Finke, and A. Waibel, “Context-dependent hybrid HME/HMM speech recognition using polyphone clustering decision trees,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 3, 1997, pp. 1759–1762.

[18] 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.

[19] J. B. Hampshire II and A. Waibel, “The Meta-Pi network: Building distributed knowledge representations for robust multisource pattern recognition,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 751–769, 1992.

[20] I. C. Jou, M. S. Hu, and Y. T. Juang, “Mandarin syllables recognition based on one class one net neural network with modified selective update algorithm,” in *Wkshp. Notes 1992 IEEE Int. Wkshp. Intell. Signal Processing Commun. Syst.*, 1992, pp. 577–591.

[21] ——, “A neural-network-based speech recognition system for isolated Cantonese syllables,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 4, pp. 3269–3272, 1997.

[22] J. F. Wang, C. H. Wu, S. H. Chang, and J. Y. Lee, “A hierarchical neural-network model based on a C/V segmentation algorithm for isolated Mandarin speech recognition,” *IEEE Trans. Signal Processing*, vol. 39, pp. 2141–2145, Sept. 1992.

[23] L. S. Lee, C. Y. Tseng, H. Y. Gu, F. H. Liu, C. H. Chang, Y. H. Lin, and Y. Lee, 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, pp. 158–179, 1993.

[24] P. C. Chang and B. H. Juang, “Discriminative training of dynamic programming based speech recognizers,” *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 135–143, 1993.

[25] 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, pp. 326–333, 1993.

[26] L. S. Lee, C. Y. Tseng, K. J. Chen, I. J. Hung, M. Y. Lee, and L. F. Chien, etc., “Golden Mandarin (II)—An improved single-chip real-time Mandarin dictation machine for Chinese language with very large vocabulary,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. II, 1993, pp. 503–506.

[27] R. Y. Lyu *et al.*, “Golden Mandarin (III)—A user-adaptive prosodic-segment-based Mandarin dictation machine for Chinese language with very large vocabulary,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. I, 1995, pp. 57–60.

[28] H. M. Wang, J. L. Shen, Y. J. Yang, C. Y. Tseng, and L. S Lee, “Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. I, pp. 61–64, 1995.

[29] 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, pt. II, pp. 206–216, Jan. 1994.

[30] C. H. Lee and B. H. Juang, “A survey on automatic speech recognition with an illustrative example on continuous speech recognition of Mandarin,” *Computat. Linguistics Chinese Language Processing*, vol. 1, no. 1, pp. 01–36, Aug. 1996.

[31] W. Y. Chen, Y. F. Liao, and S. H. Chen, “Speech recognition with hierarchical recurrent neural networks,” *Pattern Recognition*, vol. 28, no. 6, pp. 795–805, 1995.

[32] C. C. Huand and J. F. Wang *et al.*, “A Mandarin speech dictation system based on neural network and language processing model,” *IEEE Trans. Consumer Electron.*, vol. 40, no. 3, Aug. 1994.

[33] P. C. Chang, S. W. Sue, and S. H. Chen, “Mandarin tone recognition by multilayer perceptron,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 1990, pp. 517–520.

[34] Y. R. Wang, J. M. Shieh, and S. H. Chen, “Tone recognition of continuous Mandarin speech based on hidden Markov model,” *Int. J. Pattern Recog. Artific. Intell.*, vol. 8, pp. 233–246, 1994.

[35] Y. R. Wang and S. H. Chen, “Tone recognition of continuous Mandarin speech assisted with prosodic information,” *J, Accoust. Soc. Amer.*, vol. 96, no. 5, pt. 1, pp. 2637–2645, Nov. 1994.

[36] ——, “Tone recognition of continuous Mandarin speech based on neural network,” *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 146–150, Mar. 1995.

[37] B. Ma, T. Huang, B. Xu, X. Zhang, and F. Qu, “Context-dependent acoustic models for Chinese speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 1, pp. 455–458, 1996.

[38] J. L. Elman, “Finding structure in time,” *Cognitive Sci.*, vol. 14, pp. 179–211, 1990.

[39] T. W. Cacciatore and S. J. Nowlan, “Mixtures of controllers for jump linear and nonlinear plants,” in *Advances in Neural Inform. Processing Syst. 6 (NIPS’6)*, 1994, pp. 719–726.

[40] Y. Bengio and P. Frasconi, “An input output HMM architecture,” in *Advances in Neural Inform. Processing Syst. 7 (NIPS’7)*, 1995, pp. 427–434.

[41] A. Kehagial and V. Petridis, “Predictive modular neural networks for time series classification,” *Neural Networks*, vol. 10 , no. 1, pp. 31–49, 1997.

[42] S. J. Lee, K. C. Kim, H. Yoon, and J. W. Cho, “Application of fully recurrent neural networks for speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 1991, pp. 77–80.

[43] M. M. Hochberg, G. D. Cook, S. J. Renals, and A. J. Robinson, “Connectionist model combination for large vocabulary speech recognition,” In *Proc. of ICSLP-94*, 1994, pp. 1499–1502.

[44] J. D. Markel and A. H. Gray, Jr., *Linear prediction of speech*. Berlin, Germany: Springer-Verlag, 1976.

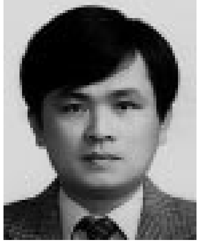
[45] H. C. Wang and H. F. Pai, “Recognition of Mandarin syllables based on the distribution of two-dimensional cepstral coefficients, *Int. J. Pattern Recognition and Artificial Intell.*, vol. 8, no. 1, pp. 247–257, 1993.

[46] Y. F. Liao, W. Y. Chen, and S. H. Chen, “Continuous Mandarin speech recognition using hierarchical recurrent neural network,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 6, 1996, pp. 3371–3374.

[47] 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. 86–90, Jan. 1998.

[48] Y. F. Liao and S. H. Chen, “An MRNN-based method for continuous Mandarin speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 2, pp. 1121–1124, 1998.

[49] T. Komori and S. Katagiri, “GPD training of dynamic programming-based speech recognizers,” *J. Acoust. Soc. Jpn. (E)*, vol. 13, no. 6, pp. 341–349, 1992.



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

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

Yuan-Fu Liao received the B.S. and M.S. degree in 1991 and 1993, respectively, and has been a Ph.D student since 1993, all in the Department of Communication Engineering, National Chiao Tung University (NCTU), Taiwan, R.O.C.

His major research interests are Mandarin speech recognition and the application of neural networks in speech processing.