Pergamon

Neural Networks, Vol. 9, No. 4, pp. 655–669, 1996

Printed in Great Britain
0893–6080/96 $15.00 + .00

0893-6080(95)00140-9

CONTRIBUTED ARTICLE

A Speech Recognition Method Based on the Sequential Multi-layer Perceptrons

WEN-YUAN CHEN,[1] SIN-HORNG CHEN[2] AND CHENG-JUNG LIN[2]

[1]Industrial Technology Research Institute, Hsinchu and [2]National Chiao Tung University, Hsinchu

(*Received* 29 *November* 1993; *revised and accepted* 7 *November* 1995)

Abstract—*A novel multi-layer perceptrons (MLP)-based speech recognition method is proposed in this study. In this method, the dynamic time warping capability of hidden Markov models (HMM) is directly combined with the discriminant based learning of MLP for the sake of employing a sequence of MLPs (SMLP) as a word recognizer. Each MLP is regarded as a state recognizer to distinguish an acoustic event. Next, the word recognizer is formed by serially cascading all state recognizers. Advantages of both HMM and MLP methods are attained in this system through training the SMLP with an algorithm which combines a dynamic programming (DP) procedure with a generalized probabilistic descent (GPD) algorithm. Additionally, two sub-syllable SMLP-based schemes are studied through application of this method toward the recognition of isolated Mandarin digits. Simulation results confirm that the performance of the method is comparable to a well modeled continuous Gaussian mixture density HMM trained with the minimum error criterion. Not only does the SMLP require less trainable parameters than the HMM system, but the former is more convenient for analysing internal features. With the aid of internal feature selection, discarding the least useful parameters of SMLP without affecting its performance is relatively easy. Copyright ©1996 Elsevier Science Ltd*

Keywords—Neural network, Generalized probabilistic descent, Multi-layer perceptrons, Hidden Markov models, Speech recognition, Dynamic programming.

1. INTRODUCTION

Speech perception in the human biological system is known to be accomplished through a network of interconnected neurons. This knowledge motivates the application of artificial neural networks (ANNs) to speech recognition because they are designed to simulate human biological neural systems. Two main approaches of ANN-based speech recognition have been studied in recent years. One is the hybrid approach which combines a conventional time-normalization procedure with a competitive neural network such as multi-layer perceptrons (MLP) (Rumelhart, 1986; Pao, 1989; Hush & Horne, 1993). A popular hybrid method employs an MLP to generate the emission probabilities of states for a continuous hidden Markov model (HMM) recognizer (Bourlard & Wellekens, 1990; Morgan & Bourlard, 1990; Renals et al., 1992, 1994; Bourlard et al., 1992). Another MLP/HMM hybrid method uses MLPs as front-end vector quantizers or labelers for a discrete HMM recognizer (Cerf et al., 1994; Rigoll, 1994). In a dynamic time-warping (DTW)/MLP hybrid method, a DTW procedure is first employed to time-align the input utterance. Next, an MLP is followed to serve as a recognizer for distinguishing time-normalized input patterns (Sakoe et al., 1989; Aikawa, 1991). Other hybrid methods which combine HMM or DTW with ANN have also been studied (Bridle, 1990; Niles & Silverman, 1990; Austin et al., 1991; Tebelskis & Waibel, 1991; Hassanein et al., 1994; Reichl et al., 1994).

Another approach is the time delay approach dealing with the time-alignment problem through mapping temporal variation of speech signals into interconnections existing between neurons of different delays (Ye et al., 1990). Time delay neural networks (TDNN) (Waibel et al., 1989; Lang & Waibel, 1990) and the temporal flow model (TFM)

Acknowledgements: The authors would like to thank Telecommunication Laboratories, MOTC, Taiwan, ROC for their support of the database. The reviewers are also appreciated for their critical comments and suggestions.

Requests for reprints should be sent to Wen-Yuan Chen, M200, CCL/ITRI, Bldg. 11, 195-11 Sec. 4, Chung Hsing Rd., Chutung, 31015 Taiwan, Republic of China (886-35-917815); E-mail: WYCHEN@M2SUN3.CCL.ITRI.ORG.TW

(Watrous, 1990) are two well-known methods of this approach. In these two methods, speech recognizers are constructed by using some basic building blocks formed by interconnecting input signals of one to three frame's delay with hidden neurons for absorbing short-time temporal distortion in the input speech signals.

A novel frame-based ANN speech recognition approach is proposed in this study. This approach directly combines the HMM method with the MLP-based pattern recognition method to employ a sequence of MLPs (SMLP) (Chen & Chen, 1991) as a word recognizer for solving the time-alignment problem. Each MLP in the SMLP is regarded as a state recognizer for distinguishing an acoustic event of the input speech signal. Next, the word recognizer is constructed through serially integrating these state recognizers. The SMLP can be made to absorb the temporal variation of speech patterns by properly controlling the time period to remain in each individual MLP. In practice, this can be simply realized by dynamic programming. Some characteristics of the proposed approach are listed as follows. First, the SMLP has a dynamic time warping capability similar to an HMM. Therefore it is suitable for the classification of dynamic speech patterns. Second, the architectural framework of the SMLP has the same topology as that of the recognized word. The former is a left-to-right MLP sequence while the latter is a left-to-right phoneme sequence. Here, a two-level competitive training algorithm is proposed for the SMLP word recognizer. Each MLP is initially trained using the well-known back propagation algorithm to distinguish the corresponding phonemes. Next, the SMLP is trained to distinguish words by a proposed word-level discriminative training algorithm which is quite different from Morgan and Boulard's work. In their hybrid HMM/MLP approach, MLP output values are considered to be estimated maximum *a posteriori* (MAP) probabilities for pattern classification. It helps frame level performance but hinders word level performance during recognition phase (Morgan & Bourlard, 1990). In contrast to the approach, the proposed word level discriminant training algorithm in this study is consistent with the recognition phase.

The rest of this paper is organized as follows. The proposed SMLP speech recognition approach is discussed in Section 2. Two sub-syllable based SMLP speech recognizers are studied in Section 3 for isolated Mandarin syllables recognition. They are based on the initial-final and the phonemic subsyllable models, respectively. Performances of these two recognizers are examined by simulations discussed in Section 4. Conclusions are finally given in the last section.

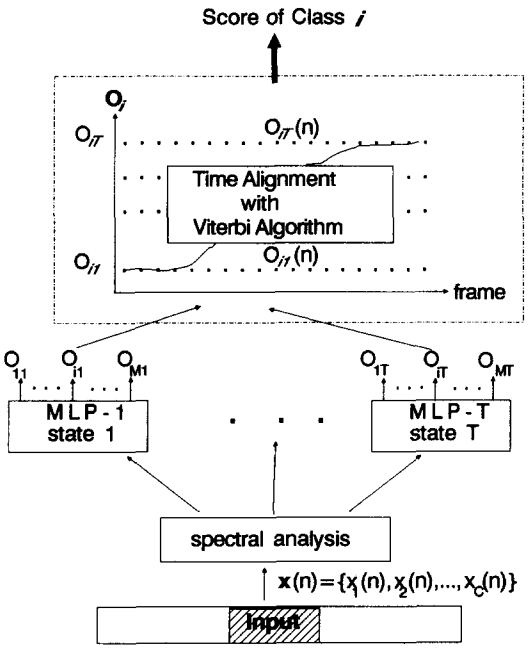

FIGURE 1. A sequential multi-layer perceptrons comprising of T MLPs. The score of the *i*th class is accumulated along the path determined by the Viterbi algorithm between the input and the *i*th class.

2. THE SMLP SPEECH RECOGNITION APPROACH

The proposed SMLP speech recognition approach is presented in this section. The basic architecture of the SMLP and the training algorithm are discussed in detail.

2.1. The Basic Architecture of the SMLP

The block diagram of an SMLP composed of T MLPs is shown in Figure 1. Each MLP is a feedforward network with M output nodes representing the M classes to be recognized. Since an MLP with two hidden layers is known to be more inclined to fall into bad local minima (Villiers & Barnard, 1993), all MLPs used in this study consist of one hidden layer only. Each MLP functions as a state recognizer for identifying an acoustic event of the input speech utterance. The operation of the SMLP is explained as follows. An input utterance $\mathbf{X}$ with N frames is taken to be a sequence of feature vectors: $X = \{\mathbf{x}(1), \ldots, \mathbf{x}(n), \ldots, \mathbf{x}(N)\}$, where $\mathbf{x}(n)$ is assumed to have C components: $\mathbf{x}(n) = \{x_1(n), x_2(n), \ldots, x_C(n)\}$. Outputs of hidden neuron j and output neuron i of the αth MLP at time n can be expressed, respectively, by

$$Y_{j\alpha}^{(\mathrm{H})}(n) = \frac{1}{1 + \mathrm{e}^{-net_{j\alpha}^{(\mathrm{H})}(n)}} \tag{1}$$

$$net_{j\alpha}^{(\mathrm{H})}(n) = \sum_k w_{jk\alpha}^{(\mathrm{H})} x_k(n) \tag{2}$$

and

$$Y_{i\alpha}^{(O)}(n) = \frac{1}{1 + e^{-net_{i\alpha}^{(O)}(n)}} \tag{3}$$

$$net_{i\alpha}^{(O)}(n) = \sum_j w_{ij\alpha}^{(O)} Y_{j\alpha}^{(H)}(n) \tag{4}$$

where weights $w_{jk\alpha}^{(H)}$ and $w_{ij\alpha}^{(O)}$, respectively, connect input node k to hidden neuron j and hidden neuron j to output neuron i in the αth MLP. By accumulating scores calculated from constituent MLPs, the discriminant function of the ith class for classification is defined as

$$g_i(\mathbf{X}, \mathbf{w}) = \left[\sum_{\theta=1}^{\Theta} S_{i\theta}(\mathbf{X}, \mathbf{w})^{\zeta}\right]^{1/\zeta} \tag{5}$$

where $S_{i\theta}(\mathbf{X}, \mathbf{w})$ is the score accumulated along the θth best path of matching $\mathbf{X}$ with the ith class, Θ is the number of warping paths, $\mathbf{w}$ is the parameter set of the SMLP, and ζ is a positive real number. The discriminant function $g_i(\mathbf{X}, \mathbf{w})$ is continuous with respect to $\mathbf{w}$ if the selection of Θ equals the total number of possible warping paths. In practical applications, only a small number of Θ best paths are evaluated due to complexity considerations. The effect on performance degradation caused by reducing Θ has been found to be insignificant (Chang & Juang, 1993). The Θ best paths search in eqn (5) can be obtained either by modifying the Viterbi algorithm to include the Θ best paths at every state where the dynamic programming procedure is performed (Schwartz & Chow, 1990), or by using the tree-trellis based fast search algorithm (Soong & Huang, 1991) which is efficient both in computation and storage. If only the best path is considered in the discriminant function defined in eqn (5), the normal Viterbi search algorithm can be directly applied to find the best path as shown in Figure 1. The $S_{i\theta}(\mathbf{X}, \mathbf{w})$ can be expressed as

$$S_{i\theta}(\mathbf{X}, \mathbf{w}) = \sum_{n=1}^{N} Y_{i\beta(n,\theta)}^{(O)}(n) \tag{6}$$

where $\beta(n, \theta)$ is the MLP corresponding to the θth warping path at the frame n, and $Y_{i\beta(n,\theta)}^{(O)}(n)$ is the value of the ith output node of the $\beta(n, \theta)$th MLP. The final decision rule involves selecting the class with a maximal discriminant function, i.e., the input utterance is recognized as the κth class if $g_\kappa(\mathbf{X}, \mathbf{w}) > g_i(\mathbf{X}, \mathbf{w})$ for all $i(\neq \kappa)$.

2.2. The Training Algorithm

In general, the error rate of a given finite set of data is a piecewise-constant function of the recognizer parameters and, thus, is not easily optimized. Juang and others (Katagiri et al., 1991; Juang & Katagiri, 1992; Chang & Juang, 1993) proposed a feasible approach to remove this difficulty. They defined a smooth 0–1 cost function to convert the mis-recognition measure into a differentiable, smooth error function to approximate the total error count. Consequently, system parameters can be optimized with respect to the smooth error function by employing gradient descent based techniques. They also developed a novel adaptive discriminant learning paradigm, i.e., the generalized probabilistic descent (GPD) algorithm, by generalizing the classical probabilistic descent method (Amari, 1967). The GPD algorithm is adopted in this study to train the SMLP.

The procedure of applying the GPD algorithm to train the weights of the SMLP is stated as follows. By using the discrimination function defined in eqn (5), the mis-classification measure for an input utterance $\mathbf{X}$ belonging to the κth class is defined as (Chang & Juang, 1993)

$$d_\kappa(\mathbf{X}, \mathbf{w}) = \frac{1}{N}\left\{-g_\kappa(\mathbf{X}, \mathbf{w}) + \left[\frac{1}{M-1}\sum_{i \neq \kappa}\{g_i(\mathbf{X}, \mathbf{w})\}^{\gamma}\right]^{1/\gamma}\right\} \tag{7}$$

where M is the number of classes and γ is a constant with a value greater than one. For simplicity, the short hand notation d for $d_\kappa(\mathbf{X}, \mathbf{w})$ is used in the following. The γ is a factor used to control the degree of participation of all competing classes in the process of optimizing the SMLP weights. In eqn (7), a negative d implies a correct classification.

The computation of eqns (5) and (7) would be quite time consuming since all of the time warping paths and competing classes are considered. One extreme case which has found extensive application in GPD studies is to let $\zeta, \gamma \to \infty$ (Komori & Katagiri, 1992; Chang & Juang, 1993). In that case, eqns (5) and (7) can be approximated by

$$g_i(\mathbf{X}, \mathbf{w}) = S_{i1}(\mathbf{X}, \mathbf{w}) \tag{8}$$

$$d = \frac{1}{N}\{-g_\kappa(\mathbf{X}, \mathbf{w}) + g_\lambda(\mathbf{X}, \mathbf{w})\} \tag{9}$$

where λ is the most probable incorrect class. Equation (8) indicates that the discriminant function is measured only along the corresponding best path ($\theta = 1$). Also, eqn (9) points out that only the correct

class and the most probable one among all incorrect classes are used in the classification decision.

Next, the SMLP parameter set, $\mathbf{w}$, is adjusted to minimize the error rate by using the GPD algorithm. A cost function, $l(d,\nu)$, is defined as (Devijver & Kittler, 1982)

$$l(d,\nu)=\int_{-\infty}^{d} h(\tau,\nu)\mathrm{d}\tau \tag{10}$$

to evaluate the cost of the current classification. Here ν is a real, positive parameter to scale the d and $h(\tau,\nu)$ is a well-behaved window function satisfying some mild conditions. When $\nu\to 0$, $h(\tau,\nu)$ is asked to converge to $\delta(\tau)$ so as to make $\ell(d,\nu)$ approximate the unit step function (i.e., $\ell(d,\nu)=1$ if $d\geqslant 0$ and $\ell(d,\nu)=0$ if $d<0$). An example of such a window function is the Gauss Laplace function, i.e.

$$h(\tau,\nu)=\frac{1}{\sqrt{2\pi\nu}}\exp\left[-\frac{1}{2}\left(\frac{\tau}{\nu}\right)^2\right]. \tag{11}$$

Notably, the cost function defined above is a monotonically increasing, differentiable function. A short-hand notation $\ell(d)$ is used for $\ell(d,\nu)$ in the following. Since a positive d implies an incorrect classification, $\sum\ell(d)$ approximately represents the total recognition error if ν approaches 0. The objective of the GPD algorithm is to recursively adjust the weights of the SMLP to minimize $\sum\ell(d)$. The change in the weights $w_{ij\alpha}^{(\mathrm{O})}$ and $w_{jk\alpha}^{(\mathrm{H})}$ can be expressed through the GPD algorithm as

$$\Delta w_{ij\alpha}^{(\mathrm{O})}=-\eta(m)\frac{\partial\ell(d)}{\partial w_{ij\alpha}^{(\mathrm{O})}} \tag{12}$$

$$\Delta w_{jk\alpha}^{(\mathrm{H})}=-\eta(m)\frac{\partial\ell(d)}{\partial w_{jk\alpha}^{(\mathrm{H})}} \tag{13}$$

where $\eta(m)$ is the learning rate at the mth iteration. The derivative terms can actually be computed through application of the chain rule as suggested by Rumelhart (1986)

$$\frac{\partial l(d)}{\partial w_{ij\alpha}^{(\mathrm{O})}}=\sum_{n|\beta(n,1)=\alpha}\frac{\partial l(d)}{\partial net_{i\beta(n,1)}^{(\mathrm{O})}(n)}\frac{\partial net_{i\beta(n,1)}^{(\mathrm{O})}(n)}{\partial w_{ij\alpha}^{(\mathrm{O})}}$$

$$=\sum_{n|\beta(n,1)=\alpha}\delta_{i\beta(n,1)}^{(\mathrm{O})}(n)Y_{j\beta(n,1)}^{(\mathrm{H})}(n) \tag{14}$$

$$\frac{\partial l(d)}{\partial w_{jk\alpha}^{(\mathrm{H})}}=\sum_{n|\beta(n,1)=\alpha}\frac{\partial l(d)}{\partial net_{j\beta(n,1)}^{(\mathrm{H})}(n)}\frac{\partial net_{j\beta(n,1)}^{(\mathrm{H})}(n)}{\partial w_{jk\alpha}^{(\mathrm{H})}}$$

$$=\sum_{n|\beta(n,1)=\alpha}\delta_{j\beta(n,1)}^{(\mathrm{H})}(n)x_k(n) \tag{15}$$

where $\delta_{i\beta(n,1)}^{(\mathrm{O})}(n)$ and $\delta_{j\beta(n,1)}^{(\mathrm{H})}(n)$ are given by

$$\delta_{i\alpha}^{(\mathrm{O})}=\frac{\partial l(d)}{\partial net_{i\alpha}^{(\mathrm{O})}(n)}=\frac{\partial l(d)}{\partial Y_{i\alpha}^{(\mathrm{O})}(n)}\frac{\partial Y_{i\alpha}^{(\mathrm{O})}(n)}{\partial net_{i\alpha}^{(\mathrm{O})}(n)}$$

$$=\frac{\partial l(d)}{\partial Y_{i\alpha}^{(\mathrm{O})}(n)}Y_{i\alpha}^{(\mathrm{O})}(n)\left(1-Y_{i\alpha}^{(\mathrm{O})}(n)\right) \tag{16}$$

$$\delta_{j\alpha}^{(\mathrm{H})}=\frac{\partial l(d)}{\partial net_{j\alpha}^{(\mathrm{H})}(n)}=\frac{\partial l(d)}{\partial Y_{j\alpha}^{(\mathrm{H})}(n)}\frac{\partial Y_{j\alpha}^{(\mathrm{H})}(n)}{\partial net_{j\alpha}^{(\mathrm{H})}(n)}$$

$$=\frac{\partial l(d)}{\partial Y_{j\alpha}^{(\mathrm{H})}(n)}Y_{j\alpha}^{(\mathrm{H})}(n)\left(1-Y_{j\alpha}^{(\mathrm{H})}(n)\right). \tag{17}$$

Next, the $\partial l(d)/\partial Y_{i\alpha}^{(\mathrm{O})}(n)$ and $\partial l(d)/\partial Y_{j\alpha}^{(\mathrm{H})}(n)$ are computed on the basis of simplified eqns (8) and (9).

We obtain

$$\frac{\partial l(d)}{\partial Y_{i\alpha}^{(\mathrm{O})}(n)}=\frac{\partial l(d)}{\partial d}\frac{\partial d}{\partial Y_{i\alpha}^{(\mathrm{O})}(n)}=\begin{cases}-\frac{1}{N}l'(d) & \text{if } i=\kappa\\ +\frac{1}{N}l'(d) & \text{if } i=\lambda\\ 0 & \text{else}\end{cases} \tag{18}$$

$$\frac{\partial l(d)}{\partial Y_{j\alpha}^{(\mathrm{H})}(n)}=\sum_{u}\frac{\partial l(d)}{\partial net_{u\alpha}^{(\mathrm{O})}(n)}\frac{\partial net_{u\alpha}^{(\mathrm{O})}(n)}{\partial Y_{j\alpha}^{(\mathrm{H})}(n)}=\sum_{u}\delta_{u\alpha}^{(\mathrm{O})}(n)w_{uj\alpha}^{(\mathrm{O})} \tag{19}$$

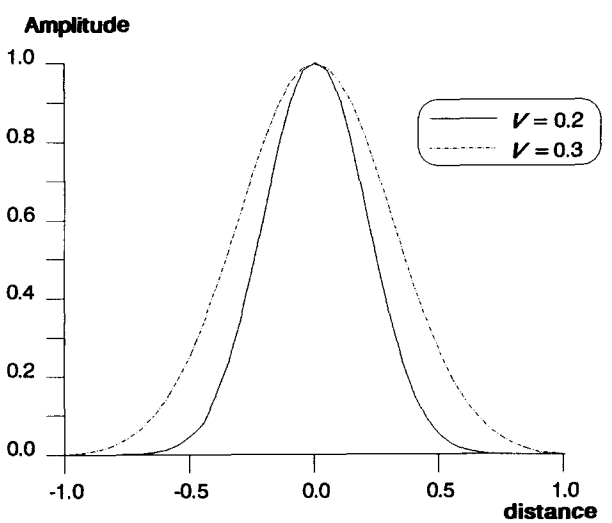

FIGURE 2. Derivative form of cost function with $\nu = 0.2$ and 0.3.

where u is an output unit and $l'(d)$ is the derivative of $l(d)$ with respect to d.

The Gauss–Laplace function as shown in eqn (11) is selected in this study for $l'(d)$. The scalar ν depends on the iteration number, i.e., $\nu = \nu(m)$ at the mth iteration. Figure 2 shows the derivative form of the cost function with $\nu = 0.3$ and 0.2. The following learning rate typically used in LVQ applications (Kohonen et al., 1988) is adopted in this study

$$\eta(m) = \eta_0\left(1 - \frac{m}{\aleph}\right) \tag{20}$$

where η_0 is a positive small number and $\aleph$ is a large positive constant. Similarly, the following scale factor $\nu(m)$ is selected, i.e.,

$$\nu(m) = \nu_0\left(1 - \frac{m}{\aleph}\right) \tag{21}$$

where ν_0 is a positive number.

The GPD algorithm discussed above is ready to train the SMLP. However, a pre-training step is added to speed up the training process due to the fact that the GPD is a rather time-consuming algorithm. Specifically, all MLPs of the SMLP are first trained independently by the error back propagation (EBP) algorithm using sub-syllable training data obtained by pre-segmenting all training utterances. Next, the GPD algorithm is applied to refine the SMLP by considering the word-level discrimination. As all MLPs of the SMLP are properly trained by the EBP algorithm, those well-recognized utterances would obtain a sizable negative measure as defined in eqn (7). Hampshire and Waibel (1990) revealed that the output state space of an MLP trained with EBP has a fraction of miss space in which utterances are mis-recognized; however, the mean square errors were still lower than those of some portions of the hit space. They demonstrated that utterances in the miss space are located near the class boundary. Therefore, the cost function defined in eqn (10) would direct the GPD algorithm to place more attention on those utterances located near the class boundary than those which are well recognized. The decreasing $\nu(m)$ would then cause the weight adjusting scheme to respond more effectively to confusing training utterances as the iteration progresses.

3. SMLP-BASED ISOLATED MANDARIN SYLLABLE RECOGNITION

In Mandarin speech, each character is pronounced as a monosyllable. An isolated Mandarin syllable can be phonetically decomposed into two sub-syllable units, i.e., initial and final. There are only 21 initials and 39 finals in Mandarin speech. The initial of a syllable may not exist and is composed of a single consonant if it exists at all. The final always exists and consists of a vowel nucleus preceded by an optional medial and followed by an optional nasal ending. The number of medials, vowels and endings in Mandarin speech are three, nine, and four, respectively. Many isolated Mandarin syllables have quite similar phoneme constituents as a result of the simple phonetic structure of the syllable. The recognition of isolated Mandarin syllables is therefore a relatively difficult task even though their size is only 408. Two SMLP-based recognition methods are proposed in this study for distinguishing Mandarin syllables. Both methods utilize sub-syllable models as basic recognition units. One uses the initial-final model while the other uses the phoneme model.

In the first method, the SMLP is composed of two MLPs which are used to discriminate initials and finals of syllables, respectively. An illustrative example of the method for recognizing isolated Mandarin digits is provided in Figure 3. In the second method, all of the Mandarin syllables are decomposed into the following phonetic structure:

$$\text{syllable} = [\text{consonant}] + [\text{medial}] + \text{vowel} + [\text{ending}]$$

where [ ] denotes an optional item. One MLP is used for each component of the above structures for phonemes discrimination. The SMLP therefore

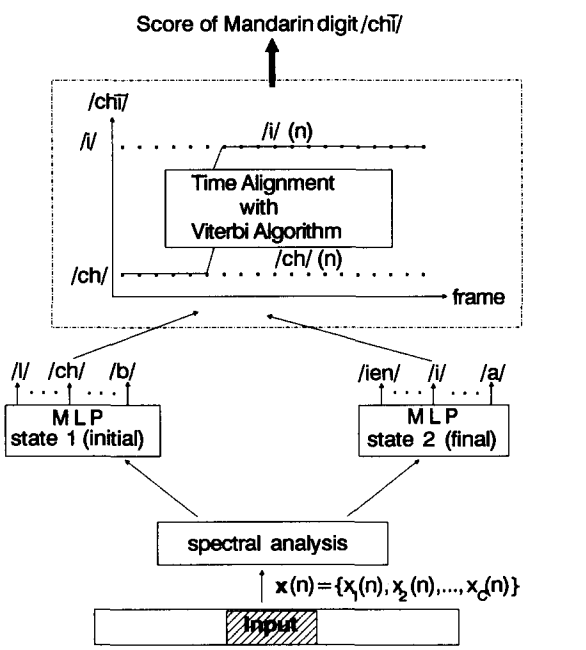

FIGURE 3. An SMLP composed of two MLPs for representing sub-syllable initials and finals. /ch/(n) and /i/(n) are respectively the outputs of /ch/ in the initial MLP and /i/ in the final MLP at time *n*.

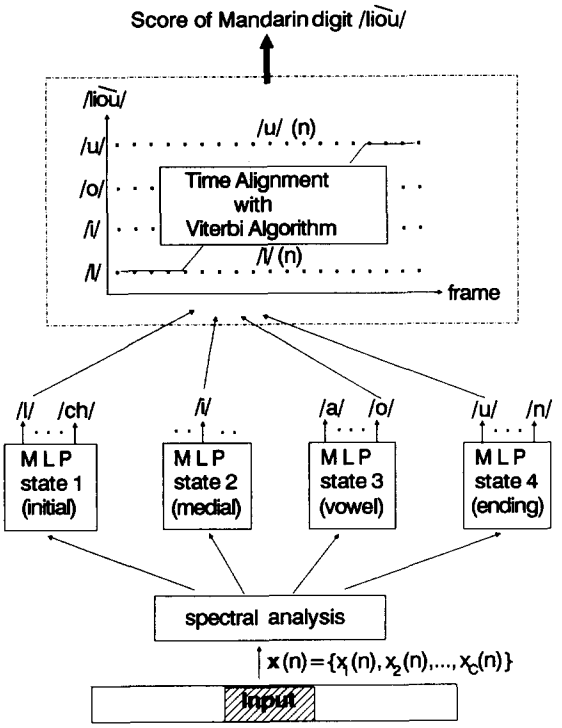

FIGURE 4. An SMLP based on the phoneme model for Mandarin syllable recognition.

consists of four MLPs in the speech recognition system. The discriminant function of a reference syllable is calculated by using only those MLPs associated with phonemes of the syllable. For instance, the second MLP is not used for the calculation of the discriminant function of the syllable /sān/ since there exists no medial phonemes in /sān/. An illustrative example of the method is shown in Figure 4.

Applying the initial-final or phonemic SMLP to isolated Mandarin syllable recognition has three distinct advantages. First, partial linguistic information that distinguishes confusing syllables has been incorporated into the architecture of the network through the use of one separate MLP for each sub-syllable unit. For instance, models are made for /l-/, /j-/ and /iou/ in the initial/final model—instead of constructing separate models for syllables /liòu/ and /jiǒu/. Complete models for these two syllables subsequently have identical second halves, thereby causing the focus of their discrimination to be shifted onto the initial components. Second, the recognition of all 408 Mandarin syllables can be decomposed into two recognition tasks, respectively, for 21 initials and 39 finals in the initial-final SMLP method, or into four recognition tasks, respectively, for 21 consonants, 3 medials, 9 vowels and 4 endings in the phonemic SMLP method. This decomposition would cause the system to be feasible because its complexity is markedly lower than with the recognition system using a single MLP. Third, a smaller training set is required for large vocabulary speech recognition since many syllables share the same sub-syllable units of initials, finals, or phonemes.

TABLE 1
Phonetic Symbols of Mandarin Digits

Digit	Yale[a]
0	lié́n[b]
1	yī
2	èr
3	sān
4	sz̀
5	wǔ
6	lioù
7	chī
8	bā
9	jiǒu

[a] The phonetic symbols are in Yale system.
[b] Tone Description
- – high level
- / high rising
- V low rising
- ⌣ high falling to low

The SMLP-based approach, although proposed here for Mandarin syllable recognition, can also be extended for speech recognition in other languages. Basically, each word can be first decomposed to recognize a concatenated sequence of phonemes, and then a phonemic SMLP is constructed to recognize words.

4. EXPERIMENTS

Although the proposed approach is potentially suitable for large-vocabulary speech recognition, its feasibility is only explored here via a simpler task of recognizing ten Mandarin digits. The phonetic structures of these ten digits are summarized in Table 1. Both recognition methods presented in Section 3 were examined. The database used in our simulations is provided by Telecommunication Laboratories (TL) (Liou et al., 1990). It consists of utterances of 100 speakers including 50 male and 50 female speakers.

TABLE 2
Distribution of Ages for the 100 Speakers

Age	15–20	21–25	26–30	31–35	35–40	41–45
Male	0	0	26	14	8	2
Female	1	4	25	15	3	2

TABLE 3
Distribution of Native Languages for the 100 Speakers

Native language	Mandarin	Amoyese	Mandarin and Amoyese	Hakkinese	Others
Male	9	32	4	4	1
Female	11	19	5	13	2

TABLE 4
Recognition Results over Testing Data of the Minimum Error Training for Continuous Gaussian Mixture Density HMM method

Input features	No. of states	No. of mixtures 2	3	4	5	6	7	8	9	10
16 energy spectra	2	89.4	92.3	94.2	94.0	93.8	94.0	94.5	93.0	93.6
	3	93.7	94.0	94.2	95.3	94.5	94.7	94.9	95.0	93.8
	4	93.8	93.6	95.1	95.0	94.6	94.0	94.4	94.3	93.3
	5	93.5	95.2	96.1	95.6	95.2	95.5	95.0	94.4	93.0
	6	94.6	95.1	96.1	95.5	96.2	95.6	94.8	95.4	93.5
	7	95.1	95.8	96.1	95.6	96.1	95.9	95.9	95.5	95.2
16 energy spectra and 16 delta energy spectra	2	94.3	94.4	95.7	97.1	97.1	97.4	97.5	97.8	97.5
	3	95.9	97.7	97.4	97.3	97.4	97.4	97.8	97.8	98.0
	4	97.4	97.7	97.7	98.1	97.8	98.2	98.0	97.7	98.1
	5	97.5	98.6	98.0	98.5	98.3	98.2	98.2	98.4	97.9
	6	98.1	98.5	98.3	98.7	98.7	98.8	98.4	98.4	98.3
	7	98.4	98.6	98.5	98.5	98.5	98.7	98.2	98.4	98.4

Each speaker repeatedly uttered the ten digits twice on different days, i.e., one repetition for training and another for testing. Notably, all the recognition results listed in this study were obtained over outside testing data. All these speakers were born and educated in Taiwan. Distributions of their ages and native languages are summarized in Tables 2 and 3, respectively. All original recordings were first collected on a SONY PCM-2500 digitization recorder through a Beyer dynamic M500N microphone in a moderately noisy room. These recordings were then played back and digitized into 16-bit samples at a rate of 20 kHz using a DSC-200 digitizer. Next, signals were pre-emphasized with a high-pass filter, $1 - 0.95z^{-1}$. A short-time spectral analysis by 512-point FFT was performed over every 25.6-ms Hamming-windowed frame with a 12.8-ms frame shift. Next, the spectrum of each frame was compressed nonlinearly into 16 triangular bands distributed in mel-scale according to a model of auditory perception (Bladon, 1985). The energy spectra of these 16 bands were then log-compressed and normalized by the average (Dautrich et al., 1983). Besides these 16 features, 16 delta energy spectra which are the difference energy spectra of two frames separated by 51.2 ms were also taken as recognition features.

Next, a series of experiments were performed on a Convex-240 parallel computer. First, a benchmark test using the continuous HMM method with a minimum error training algorithm (ME-HMM) (Rainton & Sagayama, 1992) was conducted for performance comparison. Each isolated digit was modelled in the method as a left-to-right, single-transition network. All of the HMM models were set to have the same number of states. One reasonable approach of determining the number of states in an HMM model would be to set it approximately equal to the number of phonemes of the word (Rabiner, 1989). Therefore, the optimal state number was determined here empirically by varying it from two to seven since the maximum number of phonemes in a Mandarin syllable is four. The number of the Gaussian mixture components used for every state was also varied from two up to ten to observe what accuracy was achievable when HMMs had a sufficient number of Gaussian mixture components. For minimum error training of all HMM models, the following sigmoid function was selected as the cost function:

$$l(d) = \frac{1}{1 + \exp(-\alpha(m)d)} \tag{22}$$

where

$$\alpha(m) = 1 - \frac{m}{100{,}000}. \tag{23}$$

The recognition results obtained by the ME-HMM method are listed in Table 4. The best recognition rate, 98.8%, was achieved for the case of six states and seven Gaussian mixture components with 16 spectral features and their short-term time differences as inputs.

Some parameters were determined in advance before testing the two proposed schemes. First, input recognition features of each frame were normalized to lie between -1 and $+1$ (Waibel et al., 1989). Second, both the η_0 in eqn (20) and the ν_0 in eqn (21) were empirically set to 0.3. Third, the constant $\aleph$ was set to 100,000 (= 10 digits × 100 speakers × 100 iterations).

4.1. The Initial–Final SMLP Recognition Method

From Table 1, these ten digits are composed of five initials—/l/, /s/, /ch/, /b/, /j/—and eight finals—/ien/, /i/, /er/, /an/, /z/, /u/, /iou/, /a/. As a result, the number of output nodes is set to five for the MLP representing initials and is set to eight for the MLP

TABLE 5
Recognition Results over Testing Data of SMLPs with Initial–Final Model

Input features	Condition	Initial MLP	Final MLP	Recogn. rate (%)
16 energy spectra	SMLP-A1	⟨16 30 5⟩[a]	⟨16 55 8⟩	96.4
	SMLP-A2	⟨16 35 5⟩	⟨16 60 8⟩	96.4
	SMLP-A3	⟨16 40 5	⟨16 65 8⟩	96.0
16 energy spectra and 16 delta energy spectra	SMLP-A4	⟨32 15 5⟩	⟨32 15 8⟩	98.3
	SMLP-A5	⟨32 20 5⟩	⟨32 20 8⟩	97.7
	SMLP-A6	⟨32 30 5⟩	⟨32 30 8⟩	97.7

[a] ⟨16 30 5⟩ is referred to as the MLP consisting of a two layer structure with 16 inputs, 30 hidden units and 5 outputs.

representing finals. A two-stage training procedure was applied to train the SMLP. First, the two constituent MLPs were independently trained by the conventional EBP algorithm using initial and final sub-syllable training data obtained by manually segmenting all training utterances. In the EBP training, the target was set to 0.95 for the output node of the correct class; otherwise it would be set to 0.05. After the first-stage training converges, the SMLP is then refined by the GPD training algorithm to consider word-level discrimination. Several configurations of the SMLP were examined since no relatively easy approach of determining the optimal number of hidden neurons in each MLP is currently available. Recognition results of the method are listed in Table 5. This table indicates that $\langle x, y, z \rangle$ denotes that the MLP has a two-layer structure with x inputs, y hidden units, and z outputs. For the initial–final SMLP, the best recognition rate, 98.3%, was achieved for the case of SMLP-A4.

4.2. The Phonemic SMLP Recognition Method

In the method of using phonemic SMLP, the eight finals (/ien/, /i/, /er/, /an/, /z/, /u/, /iou/, /a/) are further decomposed into one medial (/i/), seven vowels (/è/, /i/, /er/, /a/, /z/, /u/, /o/) and three endings (/ng/, /u/, /n/). The SMLP therefore has four MLPs with five, one, seven, and three output nodes, respectively. The training procedure of the method is similar to that of the initial-final SMLP recognition method. In the first-stage training, all training utterances were pre-segmented into constituent phonemes for independently training the four MLPs. This is accomplished by further dividing the final part of each utterance by a minimum error segmentation algorithm (Svendsen & Soong, 1987). Table 6 lists the recognition rates of the method for six cases which use different numbers of hidden neurons in these four MLPs. The best recognition rate, 98.8%, was achieved by the case of SMLP-B5. The performance is comparable to that of the ME-HMM method.

4.3. Discussion and Analysis

Although the best results for the ME-HMM and the SMLP achieved the same recognition rate of 98.8%, Tables 4–6 reveal that only the well modeled ME-HMMs are comparable to the phonemic SMLP. If performances of the initial–final SMLP (with two MLPs) and the phonemic SMLP (with four MLPs) are compared with those ME-HMMs having two and four states, respectively, the SMLP is superior to the ME-HMM. Besides, some advantages of the SMLP system are discussed in the following sections.

Detailed analyses of the SMLP-B5 case are worthwhile since a more thorough understanding regarding the behavior of the SMLP can be obtained. Three kinds of data analyses were performed to explore the activities of nodes in the hidden layer, i.e., the relationship between activities of hidden nodes and weights of connections to the output layer, and the segmentation of input utterances. Observations were undertaken for both well-recognized and poorly-recognized utterances.

First, the activities of nodes in the hidden layer were examined. These activities were calculated through averaging responses of overall inputs of a phoneme in a specific syllable. The activities of hidden nodes in the first MLP for those Mandarin digits having consonants are discussed here. Figure 5 illustrates the activities of hidden nodes in the first MLP. Figures 5b and 5c demonstrate that the output

TABLE 6
Recognition Results over Testing Data of SMLPs with Phoneme Model

Input features	Condition	Initial MLP	Medial MLP	Vowel MLP	Ending MLP	Rate (%)
16 energy spectra	SMLP-B1	⟨16 30 5⟩	⟨16 6 1⟩	⟨16 37 7⟩	⟨16 17 3⟩	96.1
	SMLP-B2	⟨16 35 5⟩	⟨16 8 1⟩	⟨16 42 7⟩	⟨16 20 3⟩	96.8
	SMLP-B3	⟨16 37 5	⟨16 10 1⟩	⟨16 44 7⟩	⟨16 22 3⟩	96.6
16 energy spectra and 16 delta energy spectra	SMLP-B4	⟨32 20 5⟩	⟨32 5 1⟩	⟨32 25 7⟩	⟨32 12 3⟩	98.6
	SMLP-B5	⟨32 30 5⟩	⟨32 6 1⟩	⟨32 37 7⟩	⟨32 17 3⟩	98.8
	SMLP-B6	⟨32 35 5⟩	⟨32 8 1⟩	⟨32 42 7⟩	⟨32 20 3⟩	98.5

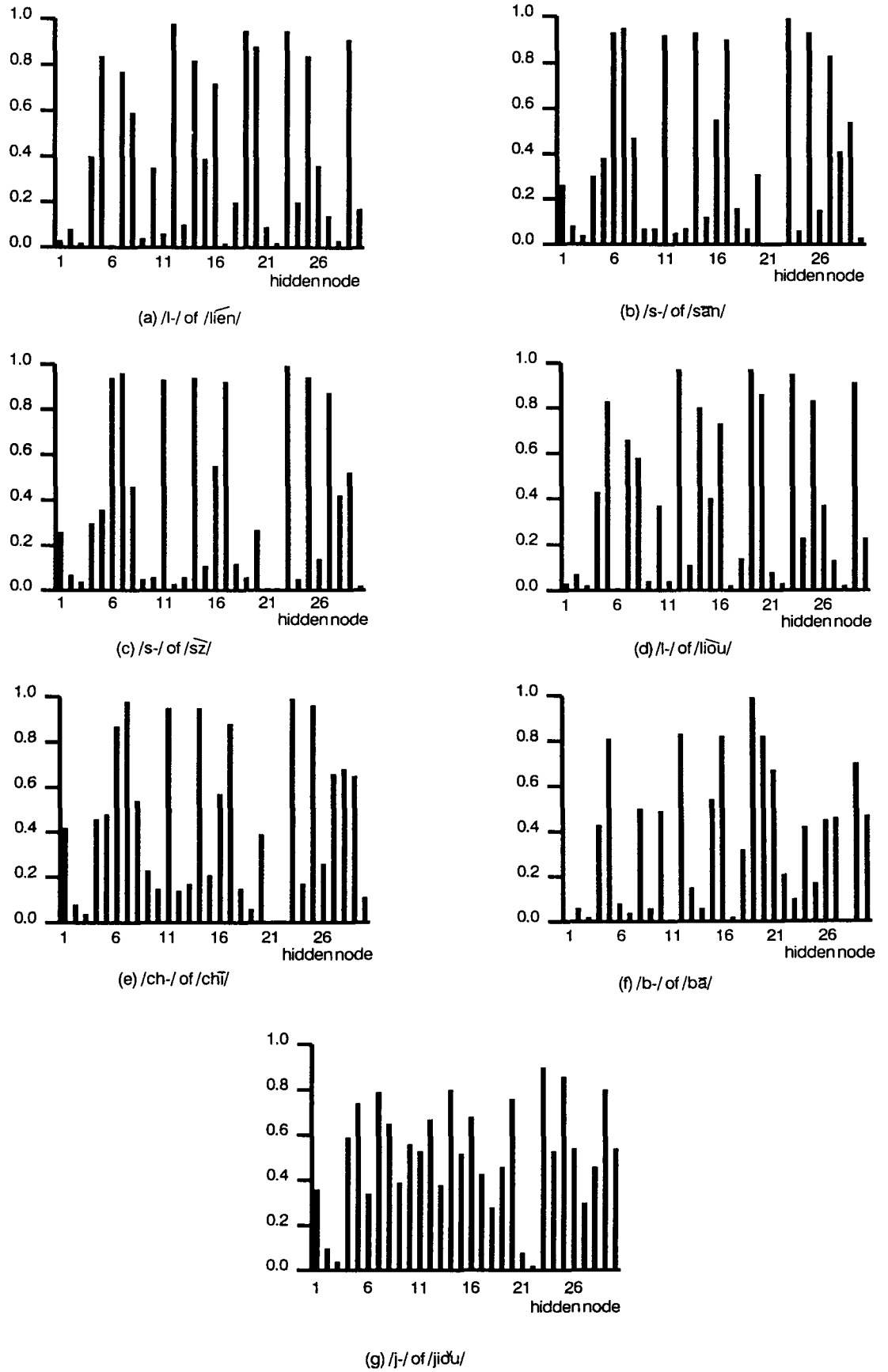

FIGURE 5. Mean of hidden outputs in the first MLP averaged overall inputs: (a) /l-/ of /lién/; (b) /s-/ of sān/; (c) /s-/ of /sż/; (d) /l-/ of /lìou/; (e) /ch-/ of /chī/; (f) /b-/ of /bā/; (g) /j-/ of /jǐǒu/.

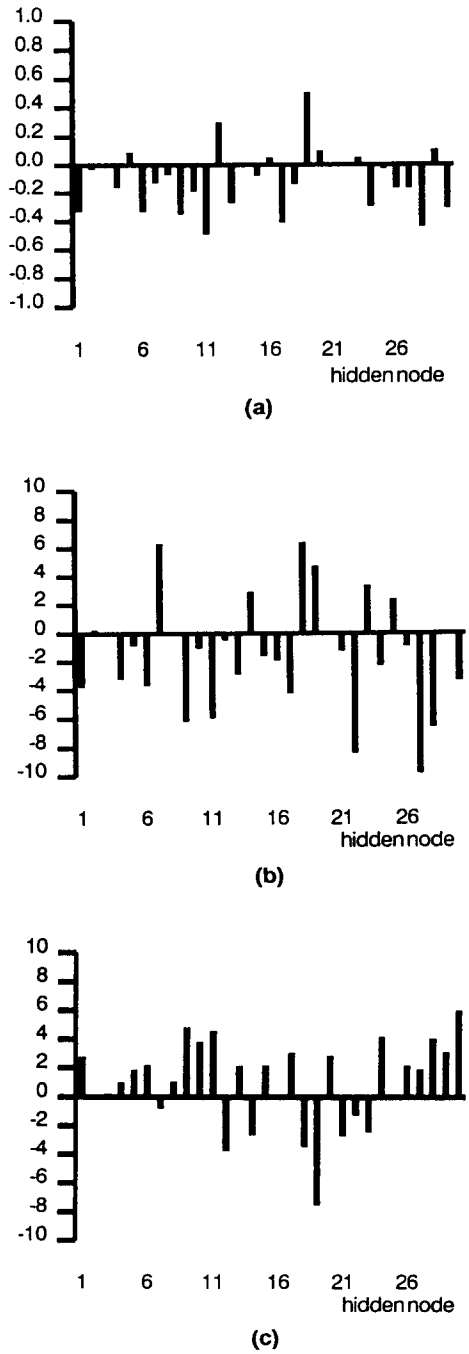

FIGURE 6. (a) Results of the acitivities of hidden nodes in Figure 5d minus those in Figure 5g; (b) weights connecting hidden nodes and output /l/ in the first MLP; (c) weights connecting hidden nodes and output /j/ in the first MLP.

patterns of hidden nodes for the inputs /s-/'s/ in /sān/ and /s ż/ are rather similar to each other in spite of the spectra of these two sounds not being exactly identical due both to different co-articulations and the tone difference in these two syllables. The same effect was observed in another illustrative example provided in Figures 5a and 5d for the inputs /l-/'s in /lién/ and /liòu/. From many observations of the same effect in the analysis, we can conclude that the

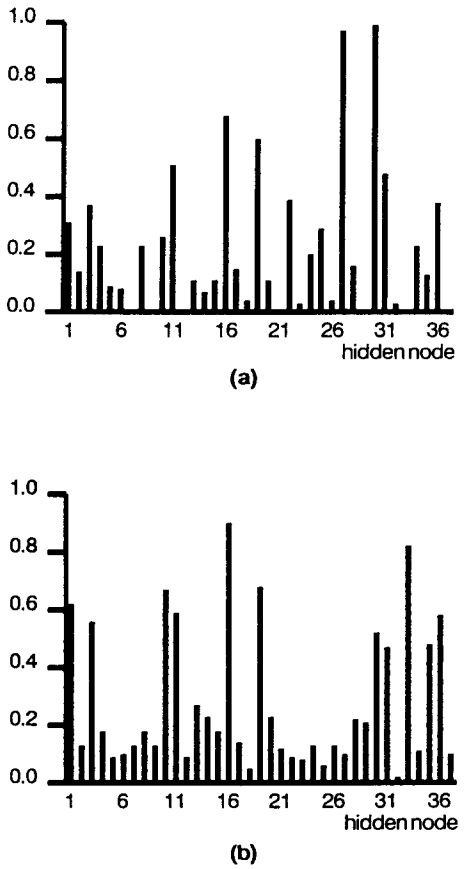

FIGURE 7. Mean of hidden node outputs in the third MLP averaged over all inputs: (a) /ī/; (b) /ŭ/.

hidden nodes of MLPs in the SMLP have learned to produce similar output patterns for inputs belonging to the same category.

The relationship between activities of hidden nodes and weights of connections to the output layer was examined next. The purpose of the analysis was to explore how the SMLP distinguishes confusing words. Therefore, only those Mandarin digits differing in a single phoneme were selected. The competition between Mandarin digits /liòu/ and /jiŏu/ was first verified. Average responses of hidden nodes in the first MLP to the input /l-/ of /liòu/ and to the input /j-/ of /jioŭ/ are calculated and displayed in Figures 5d and 5g, respectively. The differential activities of hidden nodes calculated by subtracting the average responses in Figure 5g from those in Figure 5d are displayed in Figure 6a. Weights of connections between hidden nodes and the two output nodes, /l/ and /j/, are plotted in Figures 6b and 6c, respectively. Those figures indicate that the hidden node 19, which strongly responds to the input /l-/ of /liòu/ (see Figure 6a), is positively and heavily weighted to excite the output /l/ (see Figure 6b); in addition, it is negatively weighted to inhibit the output /j/ (see Figure 6c). On the other hand, some hidden nodes, e.g., nodes 11 and 28, which strongly

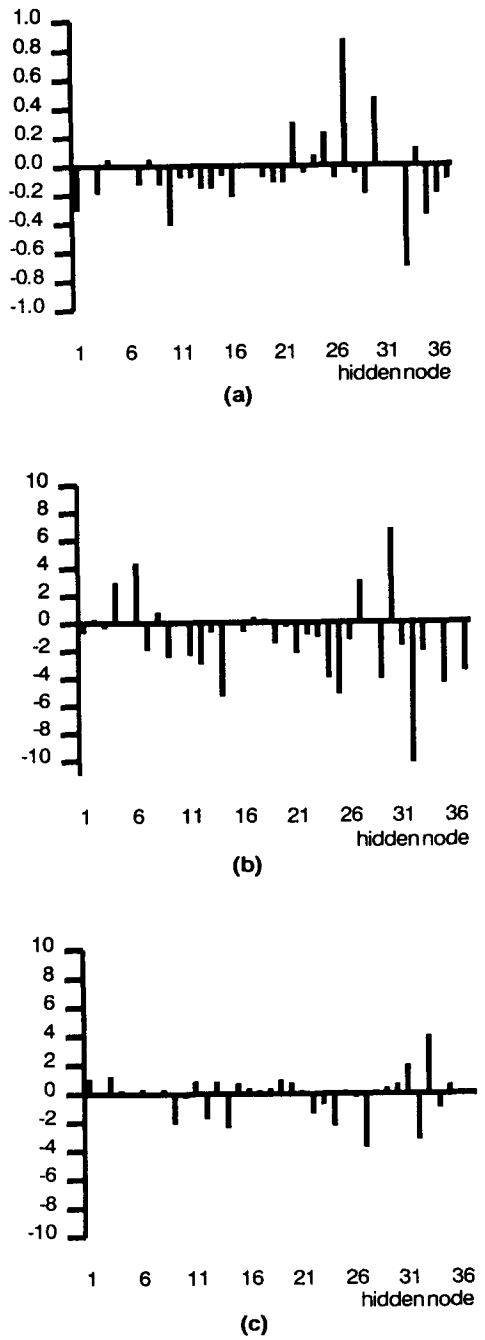

FIGURE 8. (a) Results of the activities of hidden nodes in Figure 7a minus those in Figure 7b; (b) weights connecting hidden nodes and output /i/ in the third MLP; (c) weights connecting hidden nodes and output node /u/ in the third MLP.

respond to the input /j-/ of /jiǒu/ (see Figure 6a), are positively weighted to excite the output /j/ (see Figure 6c) and are negatively weighted to inhibit the output /l/ (see Figure 6b). Another example is the competition between digits /ī/ and /ǔ/. Average responses of hidden nodes in the third MLP to inputs /ī/ and /ǔ/ are shown in Figures 7a and 7b, respectively. The differential activities of them are displayed in Figure 8a. Weights of connections between hidden nodes and the two output nodes, /i/

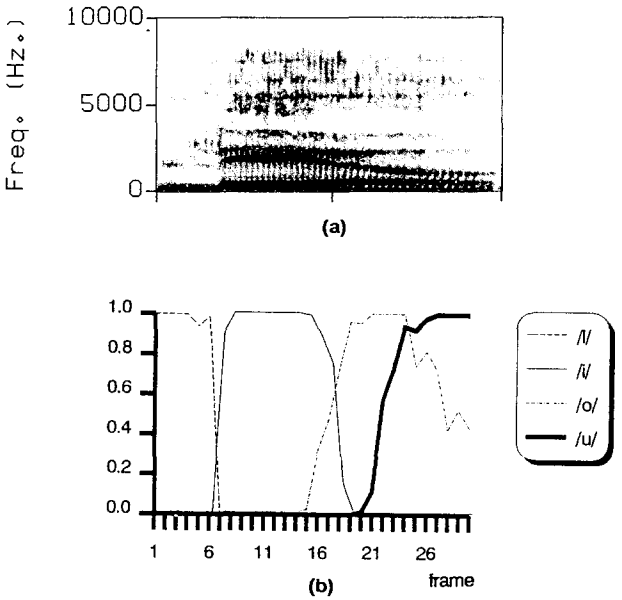

FIGURE 9. A well-recognized Mandarin digit /liòu/: (a) spectrogram; (b) outputs of MLPs corresponding to phonemes /l/, /i/, /o/ and /u/. The boundaries located at frames 7, 17, and 24 can be detected by the DP algorithm.

and /u/, are plotted in Figures 8b and 8c, respectively. These figures indicate that the hidden node 33 which strongly responds to the input /ǔ/ (see Figure 8a) is negatively weighted to inhibit the output node /ī/ (see Figure 8b) and is positively weighted to excite the output node /ǔ/ (see Figure 8c). Similar excitation and inhibition phenomena have also been observed for other pairs of confusing digits. We therefore conclude that the well-trained SMLP is highly capable of distinguishing between confusing digits.

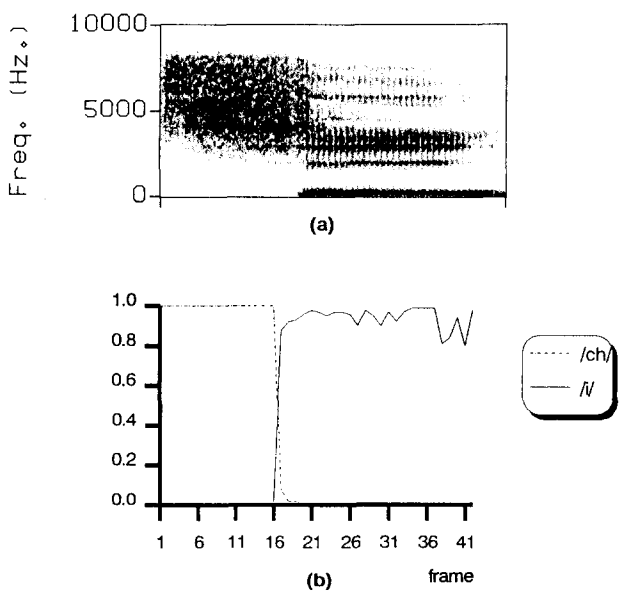

FIGURE 10. A well-recognized Mandarin digit /chī/: (a) spectrogram; (b) outputs of MLPs corresponding to phonemes /ch/ and /i/. The boundaries located at frame 17 can be detected by the DP algorithm.

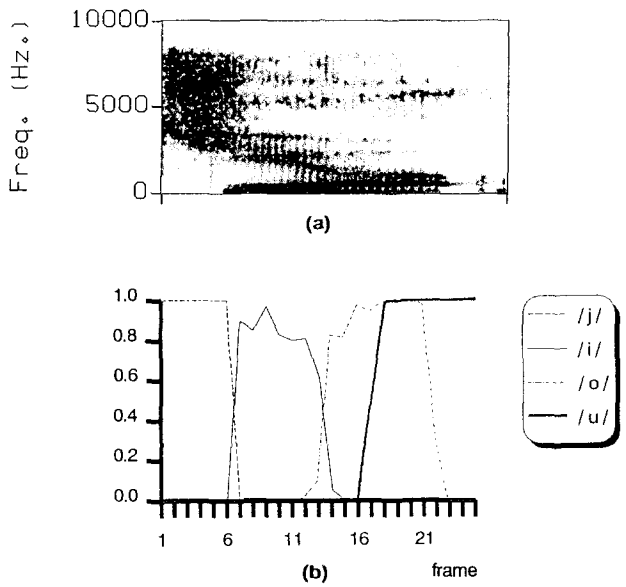

FIGURE 11. A well-recognized Mandarin digit /jiǒu/: (a) spectrogram; (b) outputs of MLPs corresponding to phonemes /j/, /i/, /o/ and /u/. The boundaries located at frames 6, 13, 18 can be detected by the DP algorithm.

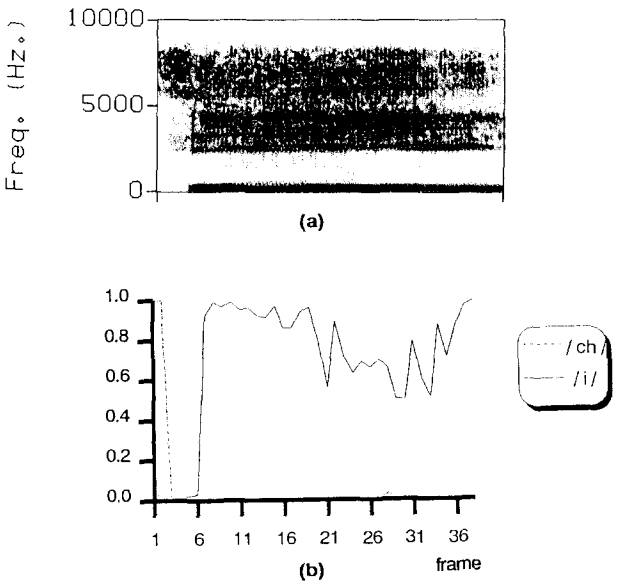

FIGURE 13. Mandarin digit /ī/ was misrecognized for /chī/: (a) spectrogram; (b) output values of MLPs corresponding to phonemes /ch/ and /ī/.

From Figures 5–8, analysing internal features is convenient for the SMLP system, but is difficult for HMM systems. Internal feature selection provides a more thorough understanding as to which parameters of the SMLP system make the greatest contribution to the recognition performance. Consequently, discarding the least useful parameters of the SMLP without affecting its performance is relatively easy. For instance, the hidden nodes 1, 2 and 23 in the first MLP are not functioning elements of the SMLP-B5 case since their outputs are very close to zero for all inputs (Figure 5). Without any re-training procedure, the recognition result over the testing data for annihilating these three hidden nodes still retains 98.8%, and 114 parameters of the SMLP-B5 have been saved.

Next, segmentations of input utterances by the SMLP were examined. Typical examples for some well-recognized utterances are shown in Figures 9–11. Spectrograms and related outputs of MLPs for these utterances are shown in part (a) and (b), respectively

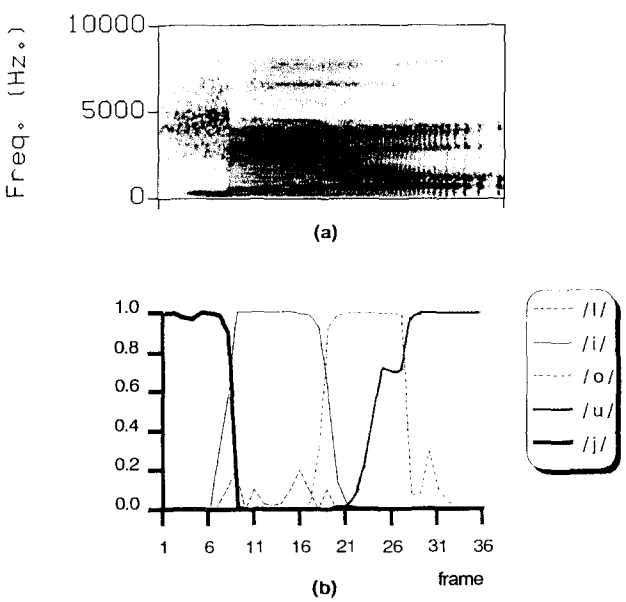

FIGURE 12. Mandarin digit /liòu/ was misrecognized for /jiǒu/: (a) spectrogram; (b) output values of MLPs corresponding to phonemes /l/, /i/, /o/, /u/ and /j/.

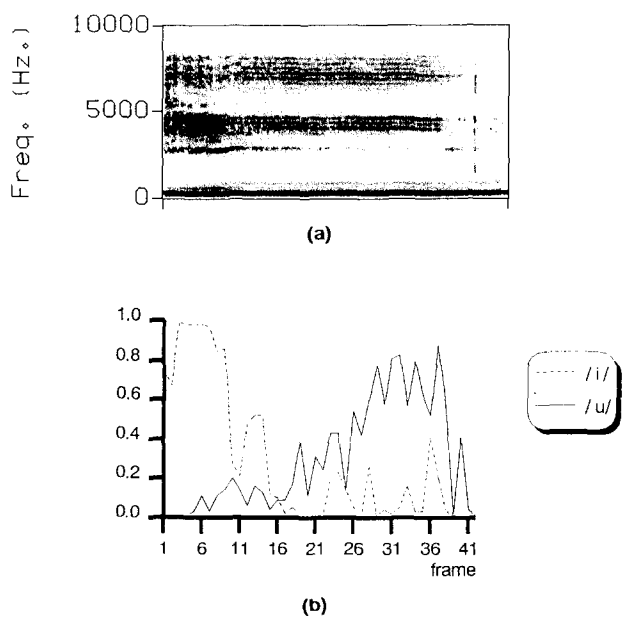

FIGURE 14. Mandarin digit /ī/ was misrecognized to /ǔ/ : (a) spectrogram; (b) output values of MLPs corresponding to phonemes /ī/ and /ǔ/ .

of these figures. Parts (a) and (b) of these figures reveal that all boundaries of phonemes detected automatically by the DP algorithm match quite well with the corresponding spectral transitions. Next, the recognition results of some incorrectly recognized utterances were examined for the sake of understanding the cause of mis-classification of the SMLP recognizer. A typical example that mis-recognizes an utterance of /liòu/ as /jiǒu/ is shown in Figure 12. Actually, only the consonant part has become confused. Figure 12a is the spectrogram of the input utterance of /liòu/. Outputs of MLPs corresponding to /l/, /j/, /i/, /o/, and /u/ are plotted in Figure 12b. This figure reveals that output /j/ attained higher values in response to the consonant part of the input utterance than those of output /l/. This would account for why the utterance was inaccurately recognized. Figure 13 shows yet another example in which an utterance of /ī/ is misrecognized as /chī/. Figure 13a reveals that most of the energy in the initial part of the utterance are located in high-frequency bands. As shown in Figure 13b, this causes the output /ch/ to strongly respond to the initial part of the input utterance and cause the mis-classification. An example of yet another type of mis-classification is provided in Figure 14. An utterance of /ī/ was inaccurately recognized as /ǔ/. Outputs on the third MLP corresponding to /i/ and /u/ are plotted in Figure 14b. This figure reveals that both /i/ and /u/ did not respond well to the entire input utterance and finally caused the recognition error. Further investigations have been performed to find the causes of those faulty segmentations. Those results indicated that most of these faulty segmentations result from the incapability of the hidden layers of the SMLP to unambiguously distinguish the correct acoustic events from incorrect ones.

Based on above analyses, we can conclude that nodes on the hidden layers of MLPs act as recognizers of basic acoustic events and the SMLP serves as a mechanism to link the sequence of detected acoustic events for forming word templates. Recognition of a syllable in the SMLP can then be regarded as distributively recognizing its own constituent acoustic events. Similar roles of hidden layers on speech recognition have also been found in TDNN by Waibel et al. (1989). Their investigation revealed that the hidden nodes on the first layer of the TDNN have learned to search for basic acoustic events; in addition, the lower layers of the network have learned to form alternate representations linking different acoustic events.

Finally, an analysis is performed of the complexities of the proposed approach and the CDHMM in terms of both the number of coefficients or weights used in their models and computations needed in the testing. Table 7 lists the numbers of coefficients used in the SMLP-B5 and the CDHMM with six states and seven mixtures. These two cases are selected because they yielded the best results in our studies of using the SMLP approach and of using the CDHMM method, respectively. This table reveals that substantially fewer coefficients were used in the SMLP-B5. The computational complexities of these two methods are analysed as follows. In the recognition phase, the main computational load is determined by calculating the model likelihoods for the CDHMM and the discriminant functions for the SMLP. Both of these two scores are computed by using the Viterbi algorithm requiring the order of $M \times K^2 \times N$ computation for the case of M classes in vocabulary size, K states in the model, and N frames of the input utterance. In this study, the best results attained when using the SMLP and the HMM are four states and six states, respectively. The computation power and the memory resource required for the SMLP system are obviously much less than the HMM system.

TABLE 7
Coefficients Used in CDHMM and SMLP for ten Mandarin Digits Recognition

CDHMM 6 states, 7 mixtures, 32 inputs	SMLP-B5
Transition prob. $6 \times 6 \times 10 = 360$	Initial MLP ⟨32 30 5⟩ 1145
Mixture coefficients $6 \times 7 \times 10 = 420$	Medial MLP ⟨32 6 1⟩ 205
Gaussian density Mean vector $6 \times 7 \times 32 \times 10 = 13,440$ Covariance matrix $6 \times 7 \times 32 \times 10 = 13,440$	Vowel MLP ⟨32 37 7⟩ 1487 Ending MLP ⟨32 17 3⟩ 615
Total = 27660	Total 3452

5. CONCLUSIONS

A novel SMLP-based approach for speech recognition has been discussed in this study. The approach can be characterized as successfully solving the time-alignment problem while retaining the competitive learning of ANN via incorporating an SMLP network with a word level discriminative training algorithm which is different from Morgan and Bourlard's work. Validation of the proposed approach has been confirmed by simulations on speech recognition of isolated Mandarin digits. The SMLP system requires less parameters and computation power than the HMM system during the recognition phase. In addition, the SMLP system provides a more feasible analysis of internal feature selection than the HMM system. Experimental results have shown that discarding the least useful parameters of SMLP through analysing the internal

feature selection without affecting the performance of the SMLP system would be relatively easy.

With its superiority in discriminating isolated Mandarin digits, future studies should extend this approach toward applications of isolated speech recognition for all Mandarin syllables.

REFERENCES

Aikawa, K. (1991). Speech recognition using time-warping neural networks. *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing* (pp. 337–346).

Amari, S. (1967). A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, **16**, 299–307.

Austin, S., Zavaliagkos, G., Makhoul, J., & Schwartz, R. (1991). A hybrid continuous speech recognition system using segmental neural nets with hidden Markov models. *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing* (pp. 347–356).

Bladon, A. (1985). Acoustic phonetics, auditory phonetics, speaker sex, and speech recognition: A thread. In F. Fallside (Ed.), *Computer speech processing* (p. 29). Englewood Cliffs, NJ: Prentice-Hall.

Bourlard, H., & Wellekens, C. J. (1990). Links between Markov models and multilayer perceptrons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, 1167–1178.

Bourlard, H., Morgan, N., Wooters, C., & Renals, S. (1992). CDNN: a context dependent neural networks for continuous speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. II-349–II-352).

Bridle, J. S. (1990). ALPHA-nets: a recurrent neural network architecture with a hidden Markov model interpretation. *Speech Communication*, **9**, 83–92.

Cerf, P. L., Ma, W., & Compernolle, D. V. (1994). Multilayer perceptrons as labelers for hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, **2**, 185–193.

Chang, P. C., & Juang, B. H. (1993). Discriminative training of dynamic programming based speech recognizers. *IEEE Transactions on Speech and Audio Processing*, **1**, 135–143.

Chen, W. Y., & Chen, S. H. (1991). Word recognition based on the combination of a sequential neural network and the GPDM discriminative training algorithm. *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing* (pp. 376–384).

Dautrich, B. A., Rabiner, L. R., & Martin, T. B. (1983). On the effects of varying filter bank parameters on isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **31**, 793–803.

Devijver, P. A., & Kittler, J. (1982). *Pattern recognition*. London: Prentice Hall International.

Hampshire, J. B., & Waibel, A. H. (1990). A novel objection function for improved phoneme recognition using time-delay neural networks. *IEEE Transactions on Neural Networks*, **1**, 216–228.

Hassanein, K., Deng, L., & Elmasry, M. I. (1994). Vowel classification using a neural predictive HMM: a discriminative training approach. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. II-665–II-668).

Hush, D. R., & Horne, B. G. (1993). Progress in supervised neural networks. *IEEE Signal Processing Magazine*, **10**, 8–39.

Juang, B. H., & Katagiri, S. (1992). Discriminative training. *The Journal of the Acoustical Society of Japan (E)*, **13**, 333–339.

Katagiri, S., Lee, C. H., & Juang, B. H. (1991). New discriminative training algorithms based on the generalized probabilistic descent method. *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing* (pp. 299–308).

Kohonen, T., Barna, G., & Chrisley, R. (1988). Statistical pattern recognition with neural networks: benchmarking studies. *Proceedings of the IEEE International Conference on Neural Networks* (pp. 61–68).

Komori, T., & Katagiri, S. (1992). GPD training of dynamic programming-based speech recognizers. *The Journal of the Acoustical Society of Japan (E)*, **13**, 341–349.

Lang, K. J., & Waibel, A. H. (1990). A time-delay neural network architecture for isolated word recognition. *Neural Networks*, **3**, 23–43.

Liou, J. S., Chen, R. G., Yu, S. M., Hwang, J. R., & Jou, I. C. (1990). The speech database of telecommunication Laboratories, Ministry of Transportation and Communications, ROC. *Proceedings of the Telecommunications Symposium, Taiwan*, (pp. 128–132).

Morgan, N., & Bourlard, H. (1990). Continuous speech recognition using multilayer perceptrons with hidden Markov models. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 413–416).

Niles, L. T., & Silverman, H. F. (1990). Combining hidden Markov model and neural networks classifiers. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 417–420).

Pao, Y. H. (1989). *Adaptive pattern recognition and neural networks*. Reading, MA: Addison-Wesley.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286.

Rainton, D., & Sagayama, S. (1992). Minimum error classification training of HMMs-implementation details and experimental results. *The Journal of the Acoustical Society of Japan (E)*, **13**, 379–388.

Reichl, W., Caspary, P., & Ruske, G. (1994). A new model-discriminant training algorithm for hybrid NN-HMM systems. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. II-677–II-680).

Renals, S., Morgan, N., Cohen, M., & Franco, H. (1992). Connectionist probability estimation in DECIPHER speech recognition system. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. I-601–I-604).

Renals, S., Morgan, N., Bourlard, H., Cohen, M., & Franco, H. (1994). Connectionist probability estimators in HMM speech recognition. *IEEE Transactions on Speech, and Signal Processing*, **2**, 161–174.

Rigoll, G. (1994). Maximum mutual information neural networks for hybrid connectionist-HMM speech recognition systems. *IEEE Transactions on Speech and Audio Processing*, **2**, 175–184.

Rumelhart, D. E. (1986). Learning internal representation by error propagation. In D. E. Rumelhart, G. E. Hinton, & R. J. Williams (Eds.), *Parallel distributed processing: exploration in the microstructure of cognition* (pp. 318–362). Cambridge, MA: MIT Press.

Sakoe, H., Isotani, R., Yoshida, K., Iso, K., & Watanabe, T. (1989). Speaker-independent word recognition using dynamic programming neural networks. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 29–32).

Schwartz, R., & Chow, Y. L. (1990). The *n*-best algorithm: an efficient and exact procedure for finding the *n* most likely sentence hypotheses. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 81–84).

Soong, F. K., & Huang, E. F. (1991). A tree trellis based fast search for finding the *n* best sentence hypotheses in continuous speech

recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 705–708).

Svendsen, T., & Soong, F. K. (1987). On the automatic segmentation of speech signals. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 77–80).

Tebelskis, J., & Waibel, A. (1991). Continuous speech recognition using linked predictive neural networks. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 61–64).

Villiers, J. D., & Barnard, E. (1993). Backpropagation neural nets with one and two hidden layers. *IEEE Transactions on Neural Networks*, **4**, 136–141.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 328–339.

Watrous, R. L. (1990). Phoneme discrimination using connectionist networks. *The Journal of the Acoustical Society of America*, **87**, 1753–1771.

Ye, H., Wang, S., & Robert, F. (1990). A pcmn neural network for isolated word recognition. *Speech Communication*, **9**, 141–153.

NOMENCLATURE

C	number of components in $\mathbf{x}(n)$
d	distance function between $\mathbf{X}$ and $\mathbf{w}$
g_κ	discriminant function of the κth class
$h(\tau, \nu)$	window function
m	iteration
M	number of classes
N	length of input utterance
$Y_{i\alpha}^{(O)}(n)$	output of output neuron i at time n of the αth MLP
$Y_{j\alpha}^{(H)}(n)$	output of hidden neuron j at time n of the αth MLP
$S_{i\theta}$	path score of the ith class along the θth best path
T	number of MLPs used in the SMLP
u	an output unit of SMLP
$\mathbf{w}$	SMLP parameter set
$w_{ij\alpha}^{(O)}$	connection weight between output neuron i and hidden neuron j in the αth MLP
$w_{jk\alpha}^{(H)}$	connection weight between hidden neuron j and input node k in the αth MLP
$\mathbf{x}(n)$	acoustic vector at time n
	$\mathbf{X} = \{\mathbf{x}(1), \ldots, \mathbf{x}(n), \ldots, \mathbf{x}(N)\}$ input utterance
$\aleph$	a large prescribed positive constant for $\eta(m)$ and $\nu(m)$
$\beta(n, \theta)$	state corresponding to the θth path at time n
$l(d, \nu)$	cost function of d with scalar ν
$l'(d)$	the derivative of $l(d)$ with respect to d
$\eta(m)$	learning rate at mth iteration
γ	constant to control degree of competing classes
κ	class corresponding to $\mathbf{X}$
λ	most probable incorrect class
$\nu(m)$	scalar at the mth iteration to control the rate of $l'(d)$
ζ	positive real number for discriminant functions
θ	warping path between $\mathbf{X}$ and class