adaptation is required to achieve performance; therefore, there is a performance versus stability tradeoff.

The fact that stability has been found to depend on the maximum rate of quantizer step size decrease is also of interest. This appears to be a justification of the shape of the quantizer scaling factor curves used in Jayant "one-word memory" type adaptive quantization.

ADPCM structures with some alternative approaches to the basic quantizer step size adaptation idea are under consideration. The theory presented here suggests that step size variation reductions will improve stability, and alternative adaptation approaches could maintain performance.

REFERENCES

[1] M. Bonnet, O. Macchi, and M. Jaidane-Saidane, "Mistracking in successive PCM/ADPCM transcoders," *IEEE Trans. Commun.*, vol. 37, Aug. 1989.

[2] ——, "Theoretical analysis of the ADPCM CCITT algorithm," *IEEE Trans. Commun.*, vol. 38, June 1990.

[3] O. Macchi and C. Uhl, "Stability of the DPCM transmission system," *IEEE Trans. Circuits Syst.*, vol. 39, Oct. 1992.

[4] R. A. Kennedy and C. R. Johnson, Jr., "Encoder stability study for the 32 kbit/s CCITT G.721 ADPCM standard based on a simplified error model," in *Proc. Second Int. Symp. Signal Processing Applications* (Gold Coast, Australia), Aug. 1990.

[5] V. Iyengar and P. Kabal, "A low delay 16 kb/s speech coder," *IEEE Trans. Signal Processing*, vol. 39, May 1991.

[6] S. Crisafulli, G. J. Rey, C. R. Johnson, and R. A. Kennedy, "A coupled approach to adpcm adaptation," *IEEE Trans. Speech Audio Processing*, pt. 1, vol. 2, no. 1, pp. 90–93, Jan. 1994.

[7] M. Honda and F. Itakura, "Bit allocation in time and frequency domains for predictive coding of speech," *IEEE Trans.Acoust., Speech, Signal Processing*, vol. ASSP-32, June 1984.

[8] P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive quantization in differential PCM coding of speech," *Bell Syst. Tech. J.*, vol. 52, Sept. 1973.

[9] N. S. Jayant, "Adaptive quantization with a one-word memory," *Bell Syst. Tech. J.*, vol. 52, pp. 1119–1144, Sept. 1973.

[10] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Clifs, NJ: Prentice-Hall, 1984.

[11] R. A. Kennedy, B. D. O. Anderson, and R. R. Bitmead, "Channels leading to rapid error recovery for decision feedback equilizers," *IEEE Trans. Commun.*, vol. 37, pp. 1126–1135, Nov. 1989.

[12] M. Vidyasagar, *Nonlinear Systems Analysis*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[13] B. D. O. Anderson *et al.*, *Stability of Adaptive Systems: Passivity and Averaging Analysis*. Cambridge, MA: MIT Press, 1986.

[14] C. R. Watkins, S. Crisafulli, and R. R. Bitmead, "An entropy coded ADPCM speech coding system for variable bit rate applications," submitted to *IEEE Trans. Speech Audio Processing*, Nov. 1994.

[15] C. R. Watkins, S. Crisafulli, R. R. Bitmead, and R. J. Orsi, "Variable bit rate ADPCM via arithmetic coding," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Adelaide, Australia), Apr. 1994.

[16] C. R. Watkins, "New techniques in signal coding," Ph.D. thesis, Australian Nat. Univ., 1994.

Generalized Minimal Distortion Segmentation for ANN-based Speech Recognition

Sin-Horng Chen and Wen-Yuan Chen

***Abstract*—A generalized minimal distortion segmentation algorithm is proposed to solve the time alignment problem for ANN-based speech recognition. By modeling dynamics of spectral information of an acoustic segment with smooth curves obtained by orthonormal polynomial expansion, a speech signal is optimally divided into segments and then recognized by an MLP recognizer. Experimental results showed that the proposed method outperforms the standard CDHMM method.**

I. INTRODUCTION

Temporal information processing is still a difficult problem to solve for artificial neural network (ANN) based speech recognition. The difficulty mainly lies in the use of fixed network structure for most existing ANN-based speech recognition systems such that the time-alignment problem is still not properly solved. One way to solve the problem is by using a hybrid approach in which a time-normalization preprocessing is attached to the neural network. The input speech signal is preprocessed to extract a fixed number of feature vectors to be fed into the neural net for recognition. Many time-normalization preprocessing methods had been studied in the past [1]–[4]. Among them, the minimal distortion segmentation (MDS) [3] is a promising method. It is based on the quasistationary assumption of the speech signal to model each utterance with a sequence of acoustic segments within which signal characteristics remain considerably uniform. The MDS algorithm is a recursive procedure that determines a set of segment boundaries such that the accumulated distortion of representing feature vectors in each segment with their average is minimum.

This correspondence introduces a new time-normalization preprocessing method for ANN-based speech recognition. It is a generalization of the conventional MDS. Instead of representing all feature vectors in a segment by their average, a set of smooth curves obtained by orthonormal polynomial expansions is used to approximate the feature contours of the segment. Because each segment usually comprises only a single acoustic event, an orthonormal polynomial expansion using a few low-order coefficients is good enough to curve fit each feature contour of a segment. The task of the generalized minimal distortion segmentation (GMDS) then becomes to find the set of boundaries that minimizes the accumulated distortion of orthonormal polynomial transforms for all segments. After determining segment boundaries, coefficients of orthonormal polynomial expansions of all segments are combined to form a feature vector for word recognition. The conventional MDS is a special case of the GMDS because only the zeroth-order coefficient is used in the orthonormal polynomial expansion. Similar work on this idea has been done on HMM's by Deng [5]. Several advantages of the proposed method can be found. First, the dynamics of the speech signal of each segment have been properly modeled in the segmentation process. Second, fewer segments or input parameters are needed to divide an utterance for

Manuscript received September 18, 1992; revised September 14, 1994. This work was supported by the National Science Council (NSC84-2213-E-009-037), Taiwan, R.O.C. The associate editor coordinating the review of this paper and approving it for publication was Dr. Xuedong Huang.

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

IEEE Log Number 9408403.

the same distortion level. Third, both static and dynamic features of spectral information for speech recognition are directly extracted in the GMDS algorithm. No resampling for feature extraction is needed [3].

II. The GMDS Algorithm

We first review the MDS algorithm proposed by Svendsen and Soong [3]. Similar work can also be found in [4]. For an input utterance $\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_T}$, the task is to divide it into m nonoverlapping quasistationary segments. Assume that the ith segment starts at frame $e_{i-1}+1$ and ends at frame e_i. Here, we simply set $e_0 = 0$ and $e_m = T$. One approach to solve this problem is to find a set of segment boundaries $\{e_1, e_2, \cdots, e_{m-1}\}$ that minimizes the following total distortion:

$$TD = \sum_{i=1}^{m} \sum_{n=e_{i-1}+1}^{e_i} d(\mathbf{X}_n, \mathbf{C}_i) \tag{1}$$

where $\mathbf{C}_i$ is the generalized center of the ith segment for the distortion measure $d(.,.)$. The MDS algorithm is a dynamic programming procedure that efficiently finds the set of the optimal segmentation boundaries. It uses the following recursive formula to calculate the optimal partial distortion $D(e_i)$ accumulated from the starting frame of the utterance to frame e_i:

$$D(e_i) = min\{D(e_{i-1}) + \sum_{n=e_{i-1}+1}^{e_i} d(\mathbf{X}_n, \mathbf{C}_i)\} \tag{2}$$

with the initial condition given as $D(e_0) = 0$. After calculating the total distortion $D(e_m)$, the optimal segmentation boundaries can be obtained by backtracking.

We now discuss the proposed GMDS algorithm. The basic idea of the algorithm is explained as follows. The speech signal is a dynamic signal in nature. As an utterance is divided into segments, it is more suitable to regard each segment as a dynamic signal rather than treating it as a static one. Therefore, instead of representing each parameter contour of a segment by a constant curve as in the case of the MDS, it is better to approximate it by a smooth curve. Distortion can hence be reduced owing to the better curve fitting. In this study, the smooth curve is an approximation of the original parameter contour obtained by an orthonormal polynomial expansion using several low-order coefficients. Specifically, let a parameter contour of a segment with length $N+1$ frames be denoted by $f(n/N), n = 0, ..., N$. The smooth curve used to approximate it can then be expressed by

$$\hat{f}(\frac{n}{N}) = \sum_{j=0}^{r} \alpha_j \phi_j(\frac{n}{N}), \qquad 0 \le n \le N \tag{3}$$

where

$$\alpha_j = \frac{1}{N+1} \sum_{n=0}^{N} f(\frac{n}{N}) \phi_j(\frac{n}{N})$$

and r is the order of the orthonormal polynomial expansion. As $r = 2$, the first three basis functions $\phi_j(\frac{n}{N})$ of the orthonormal polynomial transform can be found in [6]. The distortion between the reconstructed smooth curve $\hat{f}(n/N)$ and the original contour $f(n/N)$ is defined as

$$d(f(\frac{n}{N}), \hat{f}(\frac{n}{N})) = \sum_{n=0}^{N} (f(\frac{n}{N}) - \hat{f}(\frac{n}{N}))^2. \tag{4}$$

Assume that there are p spectral parameter contours in each segment. The total accumulated distortion for a candidate set of segmentation boundaries can be expressed as

$$TD = \sum_{i=1}^{m} \sum_{n=e_{i-1}+1}^{e_i} \sum_{k=1}^{p} d(f_k(\frac{n-e_{i-1}-1}{N_i-1}), \hat{f}_k(\frac{n-e_{i-1}-1}{N_i-1})) \tag{5}$$

where $N_i = e_i - e_{i-1}$ is the duration of the ith segment, and $f_k(\frac{n-e_{i-1}-1}{N_i-1})$ and $\hat{f}_k(\frac{n-e_{i-1}-1}{N_i-1})$ are, respectively, the kth original and the kth reconstructed spectral contours of the ith segment. Similar dynamic programming procedure like the MDS algorithm is employed in the GMDS algorithm to find the best set of segment boundaries. We note that coefficients of orthonormal polynomial expansions of feature contours in all segments of the optimal segmentation can also be obtained at the end of the GMDS algorithm. These coefficients are then combined to form an $mp(r+1)$-dimensional feature vector to be fed into the MLP for speech recognition. The block diagram of the proposed ANN-based speech recognition system is clearly shown in Fig. 1.

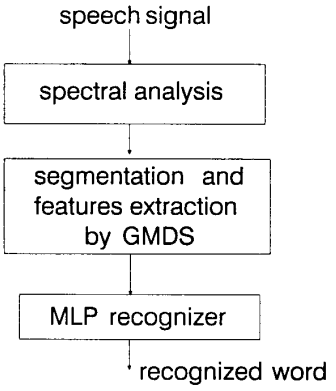


Fig. 1. Block diagram of the proposed speech recognition system.

III. Experiments

A database containing utterances of ten isolated Mandarin digits was used in the following simulations to validate the proposed recognition method. It was uttered by 100 speakers including 50 male and 50 female. Each speaker uttered these 10 digits six times on different days, with the first four repetitions being taken as the training data and the remaining two as the testing data. The phonetic symbols of these ten Mandarin digits are summarized in Table I, and the experimental condition is listed in Table II.

Taken as a reference for performance comparison, a recognition method based on the conventional continuous density hidden Markov model (CDHMM) was first tested. In this method, each isolated digit was modeled by a left-to-right network with single state transition. The number of states was varied from two to six. The observation features in each state were modeled by a five mixture Gaussian distribution. Recognition results are shown in Fig. 2. In order to distinguish between the contributions of the ANN and of the polynomial representation of the feature contour to the performance improvement achieved in the test to be discussed later by our proposed system, a recognition test by the generalized CDHMM [5] was also performed. In this approach, each feature contour of a state was described by a polynomial instead of a constant. The conventional CDHMM is a special case of the generalized CDHMM using the zeroth-order polynomial expansion. Results of the generalized CDHMM using the first- and second-order polynomial expansions are also shown in Fig. 2. Both cases of the generalized CDHMM method performed much better than the CDHMM method.

TABLE I
PHONETIC SYMBOLS OF MANDARIN DIGITS (THE PHONETIC SYMBOLS ARE IN THE YALE SYSTEM)

Digit	Yale
0	líen
1	yī
2	èr
3	sān
4	sz̀
5	wǔ
6	lìòu
7	chī
8	bā
9	jiǒu

Tone	Description
–	high level
ˊ	high rising
ˇ	low rising
ˋ	high falling to low

TABLE II
EXPERIMENTAL CONDITIONS OF THE DATABASE

DataBase	Isolated Mandarin Digits
speakers	100
number of samples	training:4000, testing:2000
sampling	20 kHz, 16 bit
preemphasized	$1\text{-}0.98z^{-1}$
acoustic analysis	512-point FFT
analysis frame	length 25.6ms, shift 12.8ms
band pass filters	16 (mel-scale)
variables/frame	16 log-compressed energy

Validation of the GMDS algorithm on segmentation and feature extraction was then investigated. Average curve-fitting distortions for various segmentations conditions are displayed in Fig. 3. We found from the figure that the average curve-fitting distortion decreases as more coefficients of orthonormal polynomial expansion are used. A typical example of segmentation and feature extraction by the GMDS algorithm is displayed in Fig. 4. Since constant curves obtained by the zeroth-order orthonormal polynomial expansions are poor approximations of the original spectral contours, it is improper to directly take their coefficients as recognition features. On the contrary, the curves obtained by orthonormal polynomial expansions using coefficients up to the second order fit well with the original spectral contours. Their coefficients are therefore suitable to be directly taken as recognition features. Because the total number of recognition features may affect the performance as well as the complexity of a recognizer, we therefore compare the average distortions of the above three cases based on the same total number of coefficients. Fig. 5 shows the relation of the average distortion to the total number of coefficients. It can be seen from the figure that the second-order orthonormal polynomial expansion still performs better.

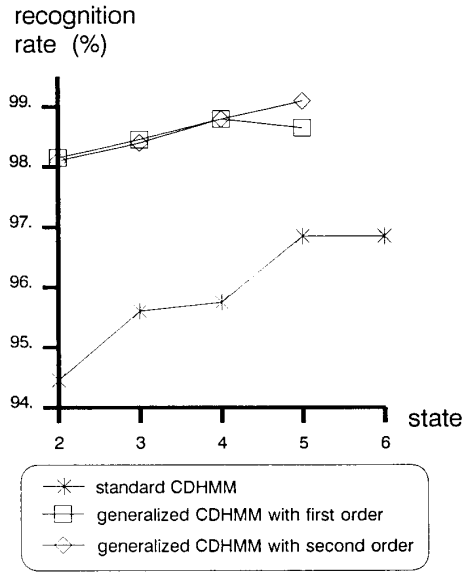


Fig. 2. Recognition results of the standard CDHMM and the generalized CDHMM. The observation features in each state are modeled by a five mixture Gaussian distribution.

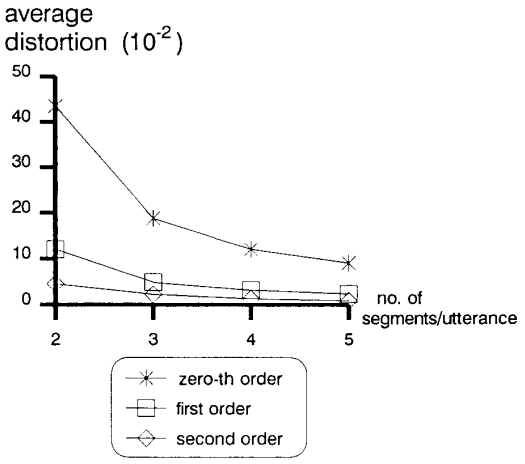


Fig. 3. Average distortion of orthonormal polynomial transforms on inside test data.

The recognition method that combines the GMDS segmentation and the MLP-based recognition was then tested. Two types of MLP's were adopted in this study. One is a three-layer MLP with two hidden layers comprising 30 and ten nodes, respectively, and another is a two-layer MLP with one hidden layer comprising 40 nodes. Both of these two MLP's were trained by the well-known backpropagration algorithm [7]. In the training, both the learning rate and the momentum factor are initially set to 0.3 and then linearly decayed as training progresses. Experimental results are shown in Fig. 6(a) for the system using three-layer MLP and in Fig. 6(b) for that using two-layer MLP. It can be seen from Fig. 6 that, for both MLP's, the cases of using the GMDS with $r = 1$ and $r = 2$ always have better recognition results than the case of using the GMDS with $r = 0$ (or the MDS). Comparing the results shown in Figs. 2 and 6, we found that the proposed approach with $r = 1$ and $r = 2$

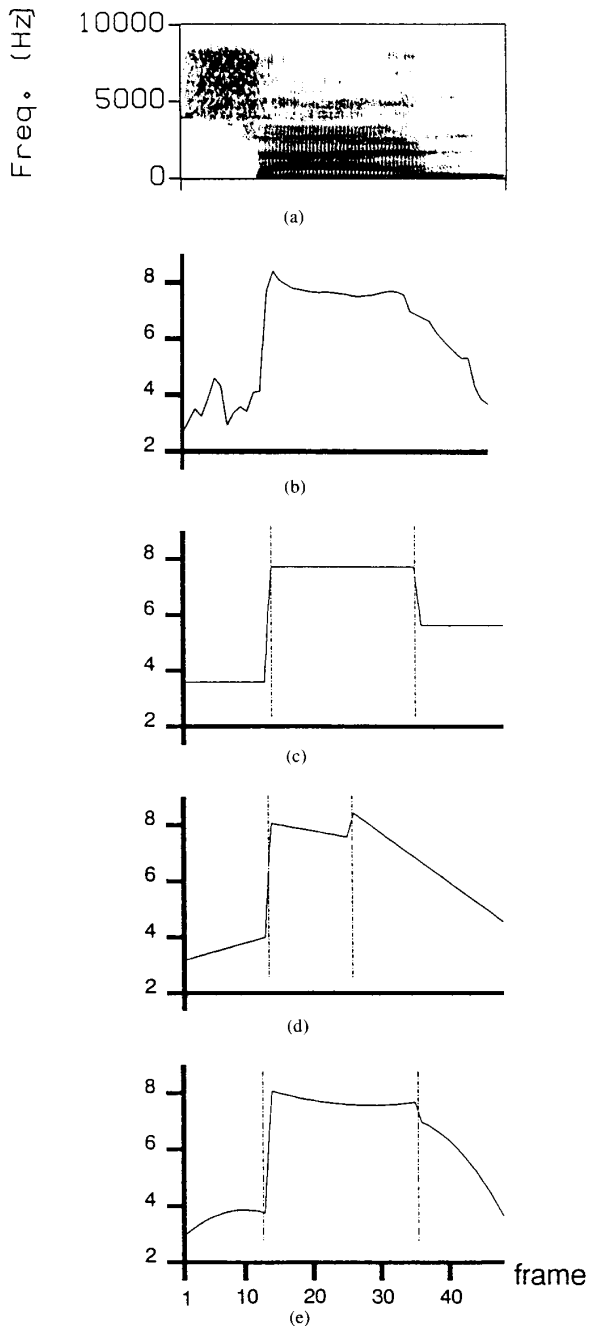


Fig. 4. Example of the GMDS algorithm. The utterance (a Mandarin syllable /sān/) is segmented into three parts: (a) Spectrogram; (b) original first band output contour. The reconstructed first band contour and the segment boundaries using (c) zeroth-, (d) first-, and (e) second-order orthonormal polynomial expansion.

always outperforms the standard CDHMM method and is comparable to the generalized CDHMM method. This shows that using segmental features, like the coefficients of orthonormal polynomial expansions used in this study, as recognition features is of great advantage in speech recognition.

In above studies, three recognition schemes including the standard CDHMM method, the generalized CDHMM method, and the MLP-based recognizer with the GMDS algorithm have been discussed. It is important to know whether or not the differences among them are statistically significant. The best results of these three schemes were therefore chosen to compare with each other using the McNemar test [8]. Suppose that the true error rates of two testing schemes are, respectively, p_1 and p_2. We would like to test the null hypothesis:

$$\mathbf{H_0}: \qquad p_1 = p_2 = p \tag{6}$$

Equations used to compute the probability p can be found in [8]. $\mathbf{H_0}$ is rejected when p is less than some significance level α (typical values of α are 0.005, 0.01, or 0.001). Results of the McNemar test on these three schemes are listed in Table III. From Table III, we have the confidence to say that the difference in performance between the MLP recognizer with the GMDS algorithm and the generalized CDHMM method is not statistically significant, and both of them perform much better than the standard CDHMM method. This also reconfirms our previous conclusion that the novel polynomial representation of feature contour makes a contribution to an improvement in recognition performance.

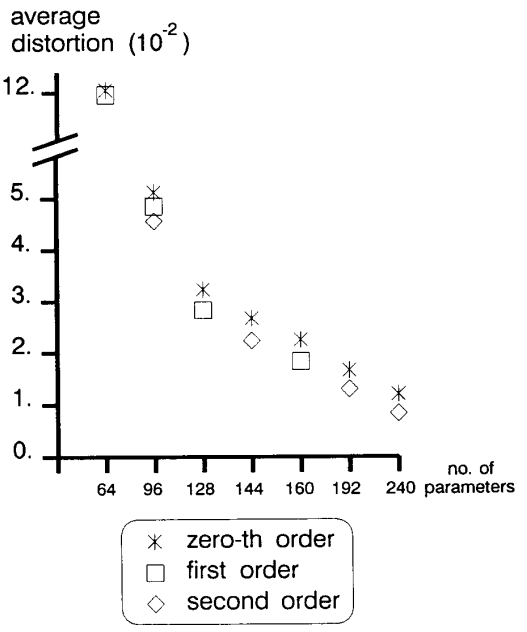


Fig. 5. Comparison of the averaged curve fitting distortion based on the same number of coefficients.

IV. Discussion

The proposed GMDS may also be applied to large-vocabulary word recognition, or continuous speech recognition by combining it with subword-based speech recognizers such as the sequential MLP recognizer [9], the segment-based DTW recognizer, etc. A possible approach is to first segment speech signals of all training utterances into phone-like subword units by the GMDS algorithm. After extracting features from these phone-like segments, the subword-based recognizer can then be properly trained. In the recognition test, a dynamic programming procedure or the level-building technique [10] can be applied to best match the testing utterance with the recognizer. An efficient recognition test can also be used by presegmenting the testing utterance using the GMDS algorithm and setting these

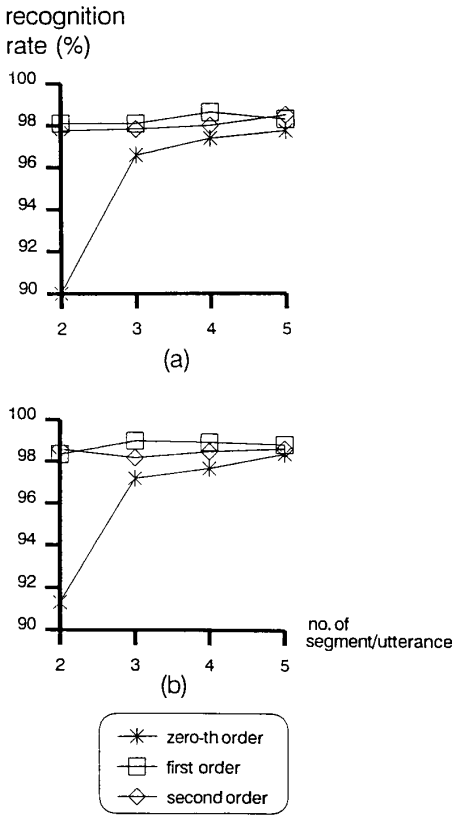
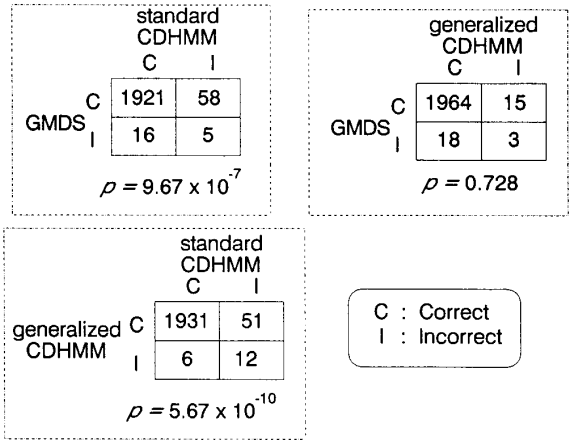


Fig. 6. Recognition results of the proposed method with varying order of orthonormal polynomial and segment: (a) Two hidden layers MLP comprising 30 and ten hidden nodes; (b) one hidden layer MLP comprising 40 hidden nodes.

TABLE III
RESULTS OF THE MCNEMAR'S TEST FOR THE STANDARD CDHMM, THE GENERALIZED CDHMM, AND THE GMDS



segmentation boundaries as candidates of subword unit boundaries. This work merits further study.

ACKNOWLEDGMENT

The authors wish to thank Telecommunication Labs, MOTC, R.O.C. for their support of the database. Thanks are also due to Mr. S. Chang for his help in getting the results of the CDHMM methods. Useful comments made by the anonymous reviewers are greatly appreciated.

REFERENCES

[1] B. R. Kammerer and W. A. Kupper, "Experiments for isolated word recognition with single- and two- layer perceptrons," *Neural Networks*, vol. 3, pp. 693-706, 1990.
[2] H. Sakoe, R. Isotani, K. Yoshida, K. Iso, and T. Watanabe, "'Speaker-independent word recognition using dynamic programming neural networks," in *Proc. ICASSP*, 1989, pp. 29-32.
[3] T. Svendsen and F. K. Soong, "On the automatic segmentation of speech signals," in *Proc. ICASSP*, 1987, pp. 77-80.
[4] Y. Shiraki and M. Honda, "LPC speech coding based on variable-length segment quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1437-1444, 1988.
[5] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, vol. 27, pp. 65-78, 1992.
[6] S. H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Trans. Commun.*, vol. 38, pp. 1317-1320, 1990.
[7] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representation by error propagation," in *Parallel Distributed Processing: Exploration in the Microstructure of Cognition.* Cambridge, MA: MIT Press, 1986, pp. 318-362, vol. 1.
[8] L. Gillick and S. J. Cox, "Some statistical issue in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, pp. 532-535.
[9] W. Y. Chen, and S. H. Chen, "Word recognition based on the combination of a sequential neural network and the GPDM discriminative training algorithm," in *Proc. IEEE Neural Networks Signal Processing*, 1991, pp. 376-384.
[10] C. S. Myers and L. R. Rabiner, "Connected digit recognition using a level building DTW algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 351-363, 1981.