JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 16, 649-660 (2000)

Short Paper

Improvement in Connected Mandarin Digit Recognition by Explicitly Modeling Coarticulatory Information

RUEY-CHING SHYU, JHING-FA WANG AND JAU-YIEN LEE*
Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan 701, R.O.C.
**Department of Electrical Engineering*
Chang Gung University
Taoyuan County, Taiwan 333, R.O.C.

The most successful training scheme for recognition of connected spoken digits is the segmental k-means algorithm, which implicitly captures the coarticulatory information of connected speech iteratively to establish reliable reference patterns. However, when this algorithm is applied to Mandarin digits, the obtained performance is inferior to that of English. Hence, a novel approach is proposed to build reliable reference patterns of connected Mandarin digits. Our method is to partition each training digit into three sections and they represent the coarticulation interacting with the preceding digit, the characteristic of the digit itself and the coarticulation interacting with the succeeding digit respectively. In this manner, the coarticulatory information is caught explicitly. Then we model these three sections separately using three Bayesian templates and a resultant multi-section Bayesian template is constructed for each reference Mandarin digit. The experimental result shows that the new method outperforms the segmental k-means by 3.2% when using a multi-speaker speech database.

Keywords: connected Mandarin digit recognition, segmental k-means algorithm, coarticulatory information, multi-section Bayesian template, level-building algorithm

1. INTRODUCTION

Connected spoken digit recognition has a wide variety of applications, such as spoken telephone dialing, utterance of an extension number by an operator to connect two telephone lines, etc. Because of the inherent context-free property of connected digits, an effective language model which can be used to correct a mis-recognized digit string can not be excessively relied on. Hence, creating a reliable digit template is the key to successful connected digit recognition. Development of the segmental k-means training algorithm [1], which implicitly captures coarticulatory information in an iterative manner, has achieved this goal; therefore, connected English digit recognition has reached a high level of performance.

Received November 30, 1998; revised April 19, 1999; accepted June 25, 1999.
Communicated by Zen Chen.

However, when this algorithm is applied to Mandarin digits, the performance obtained is inferior to that for English [2, 3]. This result can be attributed to the different language characteristics of English and Mandarin. We found that the average duration of a Mandarin digit is shorter than its English counterpart because Mandarin is a monosyllabic language. According, coarticulatory information, which should be captured by the segmental k-means iterations, is smoothed or obscured in the resulting reference pattern. To overcome this problem, we propose a novel approach which explicitly captures the coarticulatory information of connected Mandarin digits.

In our approach, each training digit is first partitioned into three sections, which represent the coarticulation interacting with the preceding digit, the characteristic of the digit itself and the coarticulation interacting with the succeeding digit, respectively. Then, these three sections (the so-called beginning, middle and end) are modeled separately using three Bayesian templates. This template is derived from Bayes' rule and trained by means of maximum-likelihood estimation. Consequently, a resultant multi-section Bayesian template is constructed for each reference Mandarin digit. By explicitly employing the coarticulatory segments, i.e., the beginning and end sections, the coarticulatory information can be protruded in the reference patterns.

In the recognition phase, the level-building algorithm [4] is employed. However, a more desirable lower and upper bound of string length for a connected Mandarin digit string is found to reduce the large number of candidate strings obtained from this algorithm [5]. In addition, a number of factors which affect recognition performance are considered to make the system more robust. Experimental results show that a string accuracy of 95.4% is achieved in multi-speaker (25 males, 25 females) mode. This paper is organized as follows. The algorithm developed to extract the coarticulatory segments is described in section 2. The formulation of multi-section Bayesian templates and the design of the corresponding training file are described in section 3. Several schemes to facilitate recognition are introduced in section 4. Section 5 gives the experimental results and discussion. Finally, conclusions are made in section 6.

2. EXPLICIT CAPTURE OF COARTICULATORY INFORMATION

The phonetic structure of a Mandarin syllable can be expressed as CV[N], where the syllable initial is a constant (C) (possibly a null), followed by a vowel (V) and an optional nasal ending [N]. Therefore, the possible cases of adjacent Mandarin digits are: (1) V|NC|V (both N, C on), (2) V|N|V (N on, C off), (3) V|C|V (N off, C on), and (4) V||V (both N, C off). Since we found that the vowel parts are less affected by the coarticulation effects [6], the contents between the two bars "| |" are regarded as the coarticulations. Furthermore, based on the inherent quasi-periodic property of the vowel parts , an automatic coarticulation segmentation algorithm can be developed as follows. First, the constituent digit boundaries of a training digit string are labeled using a fixed-length level-building algorithm [4]. Then, from each digit boundary, we search backward and forward until the pitch periods of the preceding and succeeding digits (i.e. vowel parts) are encountered. The pitch detection algorithm used is from our previous work [7]. If the number of speech samples located in this region is larger than a prescribed value, then these samples are regarded as coarticulatory information resulting from cases (1) or (2) or (3). Regarding case (4), – V||V, because the preceding and succeeding digits both exhibit periodicity, the number of speech samples located in the segmented region is rather low. Therefore, we merely take a fixed number of speech samples, in front of and behind the digit boundary, as coarticulation. We will explain the above procedure based on the example shown in Fig. 1 in the following.

The digit string shown in Fig. 1 is “8500”. It is pronounced /ba|u|ling|ling/, where /u/ is a vowel and /l/ is a periodic consonant. Using to our segmentation algorithm, the points A, B, and C are labeled first. Then, from point A, we search forward and backward; since this is the V||V case (both with periodicity), the number of speech samples located in the segmented region is rather low. Hence, a fixed number of speech samples, in front of and behind point A, are taken as coarticulation. This is the segment (a~a'). From point B, we continue to search forward and backward; since this is the V|C|V case, the forward search finds point b', and the backward search finds point b. This is the segment (b~b'). From point C, we also search forward and backward; since this is also the V|C|V case, the forward search finds point c', and the backward search finds point c. This is the segment (c~c').

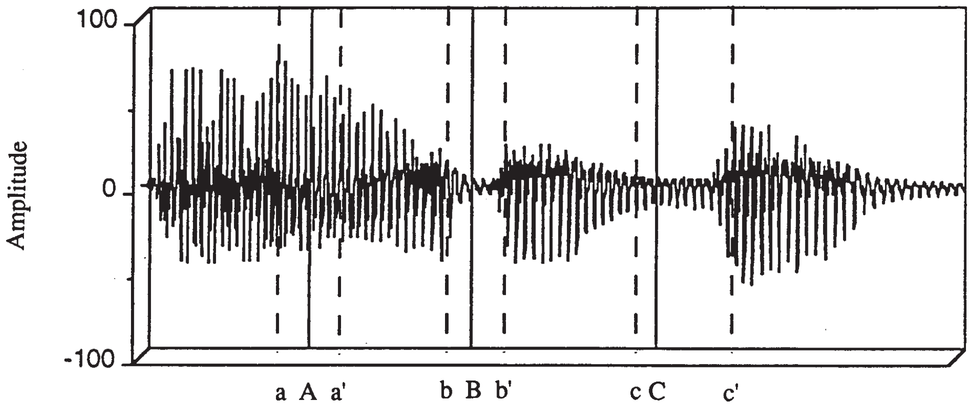

Fig. 1. Example of coarticulation segmentation. The pounts A, B and C are the word boundaries labeled by the level-building algorithm. The regions (a~a'), (b~b'), and (c~c') are the segmented coarticulations.

3. STATISTICAL MODELING OF REFERENCE PATTERNS

After the coarticulatory information at the beginning and end of a connected digit is extracted, a multi-section Bayesian template is constructed for each digit. A multi-section Bayesian template consists of three sections; the beginning, middle and end. These three sections represent the coarticulation interacting with the preceding digit, the characteristic of the digit itself and the coarticulation interacting with the succeeding digit, respectively. Each section is normalized into a fixed number of frames and organized into a Bayesian template, which is conceptually shown in Fig. 2. In this template, the feature vector of each frame is assumed to be a random sample generated from a mixed Gaussian density function. Each frame is regarded as a separate state and associated with the a *priori* probability. With respect to an input frame, each output of a Bayesian template represents the a *posteriori* probability. Defined more formally, given the *i*th input feature vector Xi, the *j*th output a *posteriori* probability of the Bayesian template Oj(Xi) is expressed as

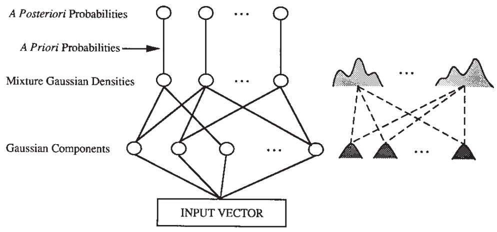

Fig. 2. Illustration of the Bayesian template in network form.

$$Oj(Xi) = \sum_{m=1}^{M_j} w_{jm} \cdot N_m(Xi; u_{jm}; U_{jm}) \cdot P_j. \tag{1}$$

Here w_m is the mixture weight of the *m*th component of a mixture Gaussian density; u and U are the mean and covariance of a Gaussian component N; M is the number of mixtures, and P is the a *priori* probability. The training approach for the Bayesian template is demonstrated below.

Training Approach for Bayesian Template

Step 0: Use the K-means algorithm to group all the feature vectors $X_1, X_2, ..., X_Q$ of a section into N clusters. Set the iteration counter J = 0 and estimate an initial Gaussian density $N_n^J(X)$, n = 1, 2, ..., N, in each cluster. Set the initial distortion $D^J = \infty$.

Step 1: Label each feature vector to identify it with the cluster it belongs to according to:

$$n^* = \arg\max_{1 \le n \le N} N_n^J(X_q), q = 1, \cdots, Q. \tag{2}$$

Step 2: Set J = J + 1 and reestimate the new feature distribution $N_n^J(X)$ using those feature vectors whose label is n.

Step 3: Calculate the distortion as

$$D^J = -\log\left(\prod_{q=1}^{Q} \max_{1 \le n \le N} N_n^J(X_q)\right). \tag{3}$$

Step 4: Check if $\dfrac{D^{j-1} - D^J}{D^J} \le \varepsilon = 0.005$ and then stop. Otherwise, go to Step 1.

After convergence, the mixture weights of each frame can be obtained. In each frame, a mixture weight is the number of training vectors which disperse into a cluster (a Gaussian component) divided by the total number of training vectors of that frame. Also, since the number of frames is fixed in a section, the a *priori* probability is the same for each frame; therefore, the a *priori* probability P in equation (1) can be omitted. The resulting multi-section Bayesian template is shown in Fig. 3.

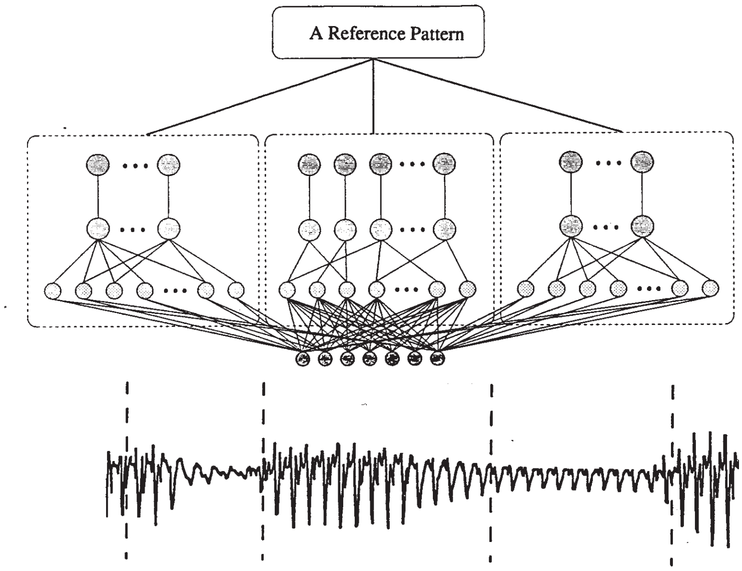

Fig. 3. The multi-section Bayesian template constructed for each Mandarin digit.

In order to coincide with the construction of multi-section Bayesian templates and to simplify the training procedure, a training file which includes all coarticulations between Mandarin digits is devised. After considering the phonemic combinations of Mandarin digits, 20 digit strings of length 4 (i.e., 60 coarticulations) are designed for the training file. These digit strings are listed in Table 1 along with the Romanization of the Mandarin digits. It is worth mentioning that all the phonemic combinations are included in this file but not all the digit combinations. Hence, some segmented coarticulations have to be shared when the multi-section Bayesian templates are trained. In addition, the phonemic combinations of /#b/ (where # represents any possible phoneme that can concatenate with /b/) are not included in this file because /b/ is a voiceless unaspirated stop which will not concatenate with the preceding phoneme.

Table 1. 20 digit strings of length 4 and the Romanization of Mandarin digits.

20 Digit Strings of Length 4 in the Training File
8969, 8719, 8500, 8301, 8202, 8103, 8051, 6721, 6522, 6231, 6325, 6159, 5749, 5547, 4345, 4077, 4142, 3739, 3529, 2709.
The Romanization of Mandarin Digits
1/i/, 2/er/, 3/san/, 4/s/, 5/u/, 6/liou/, 7/chi/, 8/ba/, 9/jiou/, 0/ling/

4. RECOGNITION APPROACHES

A block diagram of the recognition procedure of the proposed system is shown in Fig. 4. The important features are described in the following.

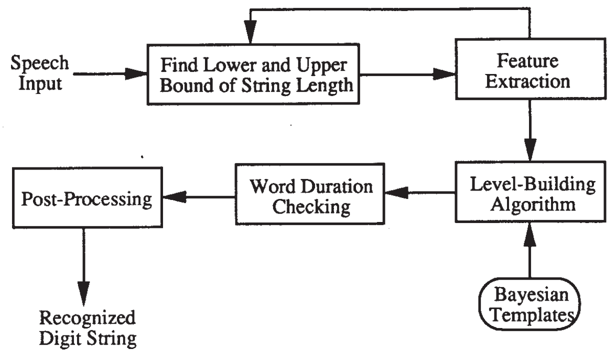

Fig. 4. Block diagram of the recognition process of the connected Mandarin digit recognizer.

4.1 Determination of the Lower and Upper Bound of the Input Digit String Length

In the recognition phase, the variable-length level-building algorithm [4] is employed. However, in order to reduce the large number of candidate strings obtained from this algorithm, a more desirable lower and upper bound of string length for a connected Mandarin digit string is searched for. Two steps are needed to determine the lower bound. In the first step, a stable-region detection algorithm is devised to detect the stable region of input speech. A stable region represents the place where the speech waveform is quasi-periodic and is composed of one or more vowels. The details of this algorithm can be found in [3]. After execting this algorithm, the lower bound of the digit string length Lmin is temporarily set to the number of stable regions found as shown in Fig. 5. In the second step, if there is more than one vowel in a stable region, different vowels can be further distinguished by the δ cepstrum [8] shown below.

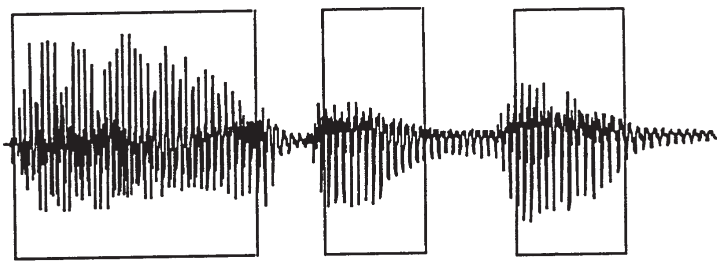

Fig. 5. Illustration of stable regions found by the stable-region detection algorithm. The number of stable regions found represents the minimum number of words in a continuous Mandarin string.

$$\delta C(i) = \sum_{n=1}^{p}\left[\sum_{k-n_0}^{n_0} kW_k C_n(i+k)\right]^2. \tag{4}$$

Here δ C(i) is the spectral transition measure at time i. $C_n(k)$ is the *n*th cepstral coefficient at time k, W_k is a window with length $2n_0 + 1$, and p is the order of cepstral analysis. If δ C(i) in a stable region is larger than a prescribed threshold, then L_{min} is increased by one. Once all the stable regions have been checked, the lower bound of the input digit string length is obtained.

Additionally, on running the stable-region detection algorithm, maximal waveform peaks appear at pitch positions. The number of peaks (called the peak count) is also recorded. Since the peaks correspond to vibrations of the vocal cord, the peak count will not change abruptly when two stable regions contain the same number of vowels. Hence, the peak counts for the stable regions with different numbers of vowels are statistically estimated. Once the distributions for different numbers of vowels are obtained, the upper bound of the string length L_{max} is determined by

$$i_k^* = \max_{N[P_k, u_i, \sigma_i] > T} (i) \tag{5}$$

$$L_{\max} = \sum_{k=1}^{Number\_Of\_Stable\_Region} i_k^* \tag{6}$$

Here N is the normal distribution, i represents the number of vowels, k is the kth stable region in the input speech, Pk is the peak count in the kth stable region and T is a threshold.

4.2 Digit Duration Checking

After candidate strings with unreasonable lengths are removed using Lmin and L_{max}, the remaining candidate strings are checked based on their durations to choose the most likely result. For Mandarin digits, we have found that, if the number of digits in a continuous digit string is fixed, the average digit length will not vary too much. Hence, a suitable digit duration is set from $\frac{\text{Total Frames}}{\text{Number of Digits}} \times 0.8$ to $\frac{\text{Total Frames}}{\text{Number of Digits}} \times 1.2$. (Here, the total number of frames includes the original frames, not the normalized frames.) The remaining candidate strings are first ordered according to their accumulated distances obtained from the level-building algorithm. If the digit duration of a candidate string with a minimum accumulated distance fails the above criterion, then the next candidate string is checked. In this manner, the chosen string has the minimum accumulated distance among the candidate strings whose durations are reasonable.

4.3 Post-Processing of Confusing Pairs

Confusing pairs (1/i/, 7/chi/) and (6/liou/, 9/jiou/) in Mandarin digits cause a number of substitution errors; hence, these pairs are post-processed. In the stable-region detection algorithm, the consonant parts of continuous Mandarin speech can also be labeled [3]. By emphasizing the consonant parts of these pairs, substitution errors can be effectively reduced. To meet this requirement, a weighting function shown in Fig. 6 is used based on our earlier study [9]. The confusing pairs are re-recognized using the DTW method. The optimum path is first evaluated, and each local distance along this path is multiplied by the weighting function. After weighting is conducted, each element in a confusing pair will have a new accumulated distance. The smaller one is chosen as the final result.

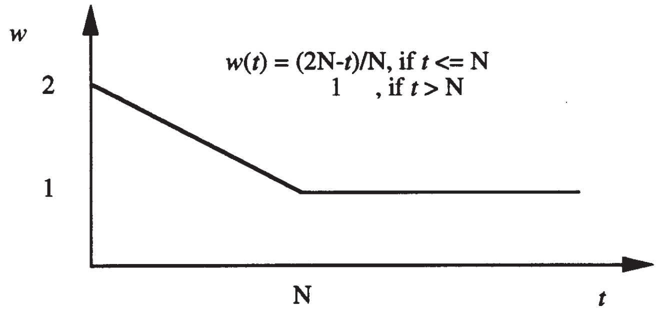

Fig. 6. Angle weighting function used to emphasize the consonant part of a Mandarin digit. N is the consonant position labeled by the stable-region detection algorithm.

4.4 Computational Considerations

When the level-building algorithm is executed, the most time-consuming operations are the calculations of local distances. The calculated local distance is the negative log of the output from the Bayesian template and has the form: $-\log\left(\sum_{m=1}^{M} w_m O_m(X)\right)$, where M is the number of mixture components, w_m is the weight and $O_m(X)$ is the Gaussian density function for the mth mixture component. Obviously, this equation requires considerable computation, especially the log operation outside the parentheses and the exponential operation inside $O_m(X)$. However, if the mixture Gaussian density of the above equation, $\sum_{m=1}^{M} w_m O_m(X)$, is replaced by the partitioned Gaussian regressive mixtures [10], $\max_{1\le m\le M} w_m O_m(X)$, the computational load can be alleviated because the operations of log and max can now be interchanged. As a result, the log operation and the terms, including wm and some parameters of $O_m(X)$, can be combined into a constant because these terms are known after training. In addition, the exponential operation inside $O_m(X)$ can also be eliminated by the log operation.

5. EXPERIMENTAL RESULTS

The speech data was provided by 50 inexperienced speakers (25 males and 25 females). Each speaker spoke the isolated digits and digit strings in the training file three times for training purposes. In addition, another 100 digit strings with lengths varying from 2 to 7 were spoken twice for testing purposes [11]. Speech data was digitized at a sampling rate of 8 kHz and pre-emphasized. Speech samples were blocked into a series of frames, and each frame was multiplied by a Hamming window. The LPC derived cepstral vector was computed up to the 12-th component. Bandpass filtering [12] was then applied to the cepstral coefficients. Each frame was computed with a 24 dimensional feature vector of 12 cepstral coefficients and 12 delta cepstral coefficients.

5.1 Effect of the Number of Frames in the Beginning, Middle, and End Sections

Since each section of the multi-section Bayesian template is normalized into a fixed frame, different numbers of frames were tried to choose the optimal values. The results are given in Table 2. The best performance was found at the 4, 8 and 4 frames of the beginning, middle and end sections, respectively. This result indicates that coarticulation indeed plays an important role in creating a robust connected digit template. In addition, the optimal proportion of these three sections was 1:2:1.

Table 2. Recognition performances among different numbers of frames of the beginning, middle and end sections.

Number of Frames of Middle Section / Number of Frames of Beginning and End Section	4	6	8	10
2	86.1%	87.4%	88.2%	88.3%
4	90.4%	90.2%	92.7%	91.6%

5.2 Effect of Using Bayesian Templates

When Bayesian templates were used to represent reference patterns, multi-templates per digit and the number of mixtures were used to represent the frame density. The number of templates per digit was set at 3, 6, 9, and 12. The number of mixtures used to represent the frame density was set at 5, 7, and 9. The results are shown in Table 3 and presented graphically in Fig. 7. The best performance occurred at 9 templates/digit with 7 mixtures/frame. The poorer performance obtained in the other cases was due to either a lack of sufficient data for training the Bayesian templates or to the fact that the acoustic resolution of the Bayesian template was not fine enough.

Table 3. Recognition rates among different cases of templates/digit and mixtures/frame of the Bayesian template.

Templates/Digit / Mixtures/Frame	3	6	9	12
5	85.3%	88.2%	91.4%	92.5%
7	86.1%	89.6%	92.7%	92.1%
9	85.1%	91.3%	92.2%	91.8%

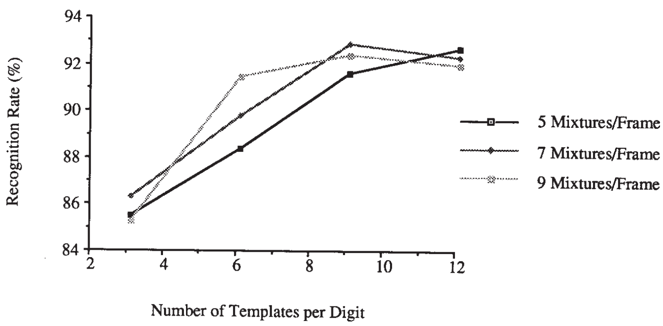

Fig. 7. Recognition rates among different cases of templates/digit and mixtures/frame of the Bayesian template.

5.3 Effect of Finding the Lower and Upper Bound

If the lower and upper bound of the input string length are not set, the level-building algorithm will stretch to the prescribed maximum length (i.e., 7 in our database). In addition, the level-building algorithm provides multiple candidates at each level. Hence, R^L candidates strings of length L are provided for an unknown input string, where R is the number of candidates per level (typically R=2) [5]. In this manner, the number of candidate strings for an unknown input string is $2^1 + 2^2 + ... + 2^7 = 254$. In our experiment, after the lower and upper bound were found, the average number of candidate strings was reduced to about 76, which is a 70% reduction.

5.4 Effect of Digit Duration Checking

The performance with and without digit duration checking is shown in Table 4. It can be seen that digit duration checking assisted in removing insertion and deletion errors. Insertion and deletion errors were reduced by 0.6% and 0.8%, respectively. This improvement is not trivial since the testing database used was a multi-speaker one.

Table 4. Performance comparison between word duration checking and no word duration checking.

Word Duration Checking	Insertion Errors	Deletion Errors
No	1.6%	2.6%
Yes	1.0%	1.8%

5.5 Effect of Post-Processing

The performance before and after post-processing is shown in Table 5. The results show that the substitution errors were decreased by 1.3%. This improvement was not as large as we expected. One possible reason may have been that the consonant parts of connected Mandarin speech are too short; consequently, consonant parts are still obscured by the vowel parts in distance accumulation even when they have been emphasized. After post-processing, we were able to achieve a string accuracy of 95.4%.

Table 5. Performance comparison between before post-processing and after post-processing.

Post-Processing for Confusing Pairs	Substitution Errors
No	3.1%
Yes	1.8%

6. CONCLUSIONS

A robust connected Mandarin digit recognition system based on multi-section Bayesian templates and the level-building algorithm has been presented in this paper. Coarticulatory segments are used explicitly to establish the multi-section Bayesian template such that both coarticulation effects and the characteristics of the digit itself can be properly exhibited. In particular, each section is organized into a Bayesian template which is characterized by a powerful statistical framework. In the recognition phase, several tasks, such as finding a more desirable lower and upper bound of the input digit string length, digit duration checking and post-processing, are carefully executed to make the proposed system robust. The proposed system has been tested using a multi-speaker database (25 males, 25 females), and a string accuracy of 95.4% was achieved.

REFERENCES

1. L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A segmental k-means training procedure for connected word recognition," *AT &T Tech. Journal*, Vol. 65, No. 3, 1986, pp. 21-31.
2. C. H. Lin, E. F. Huang, and T. H. Chiang, "The recognition of connected Mandarin digits based on continuous hidden Markov model," Telecommunication Conference, 1990, pp. 144-149.
3. R. C. Shyu, C. C. Huang, J. F. Wang, C. H. Wu, J. S. Shyuu, and J. Y. Lee, "Several specially designed approaches for connected Mandarin digit recognition," *International Computer Symposium*, 1992, pp.479-485.
4. C. S. Myers and L. R. Rabiner, "Connected digit recognition using a level-building DTW algorithm," *IEEE Transactions on Acoustics Speech and Signal Processing*, Vol. 29, 1981, pp. 351-363.
5. L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High performance connected digit recognition using hidden Markov models," *IEEE Transactions on Acoustics Speech and Signal Processing*, Vol. 37, 1989, pp. 1214-1225.
6. R. C. Shyu, J. F. Wang, C. C. Huang, C. H. Wu, and J. Y. Lee, "A vowel-driven connected Mandarin digit recognition system," *Journal of Information Science and Engineering*, Vol. 12, 1996, pp. 365-379.
7. 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 Transactions on Signal Processing*, Vol. 39, No. 9, 1991, pp. 2141-2146.
8. T. Svendsen and F. K. Soong, "On the automatic segmentation of speech signals," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 1987, pp.77-80.
9. R. C. Shyu, J. F. Wang, C. H. Wu, M. Y. Chen, and J. Y. Lee, "A robust connected Mandarin digit recognizer based on Bayesian template," *Computer Processing of Chinese and Oriental Languages*, Vol. 7, 1993, pp. 45-60.
10. B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Transactions on on Acoustics Speech and Signal Processing*, Vol. 33, 1985, pp. 1404-1413.
11. C. J. Huang, P. Y. Ting, T. H. Chiang, and E. F. Huang, "Performance comparisons among different connected word recognition algorithms using Mandarin digits," *TI Technical Journal (ROC)*, Vol. 19, No. 3, 1989, pp. 303-312.
12. B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Transavtions on Acoustics Speech and Signal Processing*, Vol. 35, 1987, pp. 947-953.

Ruey-Ching Shyu (®}· ç¼y) was born in Hsinchu, Taiwan, on November 2, 1964. He received the B.S. and M.S. degrees in Electrical Engineering from National Cheng Kung University in 1987 and 1989, respectively. He is now working towards the Ph.D. degree in the same department. His research interests include speech recognition, digit signal processing, and neural networks.

Jhing-Fa Wang (¤ýÂ@µo) is now a professor at National Cheng Kung University, Tainan, Taiwan. He received his B.S. and M.S. degrees in Electrical Engineering from National Cheng Kung University in 1979 and 1973, respectively, and the Ph.D. degree in the Department of Computer Science & Electrical Engineering from the Stevens Institute of Technology, U.S.A., in1983. He is now also an associate editor of IEEE Transactions on Neural Networks and VLSI systems. He received outstanding achievement awards from the Institute of Information Industry in 1991 and the National Science Council in 1990, 1995 and 1997, respectively. His current research areas include VLSI/CAD, speech recognition, speech coding, optical character recognition and natural language processing. He was elected as an IEEE Fellow in 1999 for contributions to software-hardware co-development of large-vocabulary Mandarin speech processing and recognition systems.

Jau-Yien Lee (§õ»FÄY) received the B.S., M.S., and Ph.D. degrees in EE from National Cheng Kung University, Taiwan, in 1966, 1970 and 1974, respectively. From 1970 to 1982, he served as a faculty member and later as Chairman of the Department of the EE in the Military Academy, ROC. In 1982, he retired from the Army and joined NCKU, where he was a professor. One year later, he was elected Chairman of the department. During 1982-1993, he set up a VLSI/CAD and communications laboratories to promote VLSI/Comm integration research, and during that time he published more than 100 journal and conference papers. In 1993, he joined Chang-Gung University as Chairman of the E.E. Department and later as Dean of Engineering. In 1999, he retired from CGU and was invited to serve as Chair professor of the University.