Isolated Mandarin Base-Syllable Recognition Based upon the Segmental Probability Model

Ren-Yuan Lyu, I-Chung Hong, Jia-Lin Shen,
Ming-Yu Lee, and Lin-Shan Lee

***Abstract*—In this correspondence, a segmental probability model (SPM) is proposed for fast and accurate recognition of the highly confusing isolated Mandarin base-syllables by deleting the state transition probabilities of continuous density hidden Markov models (CHMM), abandoning the dynamic programming process, letting the states equally segment the base-syllables deterministically, and using several special approaches to improve the accuracy and speed. This is achieved by considering the special characteristics of the target vocabulary.**

***Index Terms*—Hidden Markov model, isolated syllable, Mandarin Chinese, segmental probability model, speech recognition.**

I. Introduction

THE recognition of the 416 Mandarin base-syllables (i.e., syllables disregarding the tones) is a key problem for Mandarin speech recognition with very large vocabulary [1]–[3]. However, this is in fact a very difficult task, because there exist 38 confusing subsets in this set of base-syllables. A good example of such confusing subset is the a-set, {ba, pa, ma, fa, da, ta, na, la, ga, ka, ha, ja, cha, sha, dsa, tsa, sa, a}, among which the only discriminative speech segment is in the initial consonant part of each syllable with very short duration [4]. Specially trained continuous density Hidden Markov Models (CHMM's) have been used in previous studies, and high recognition rates on the order of 93.89% [5], [6] have been achieved for speaker-dependent and isolated-syllabic mode, which is also the case considered in this work. Although those CHMM-based approaches have achieved very successful recognition rates, they suffered from not only very high computational load in both training and recognition phases, but also the time-consuming process of human-aided segmentation of the training data for the discriminative initial consonant part of each syllable. To alleviate those problems, a new approach specially developed for isolated Mandarin base-syllable recognition, referred to as the *segmental probability model* (SPM) [2] is proposed in this work. This approach can be viewed as a modified version of CHMM with the state transition probability matrix abandoned and the speech utterances equally segmented by the states. Consider the fact that isolated Mandarin base-syllables have relatively simple phonetic structures and the fact that the primary purpose in such a recognition task is distinguishing each base-syllable from the others instead of decoding it into a few phonemes with their boundaries. It is therefore reasonable to assume that there is no need for the optimal sequence decoding by a dynamic programming procedure, usually performed in traditional CHMM approaches. The computational load can thus be reduced significantly both in recognition and training.

Manuscript received February 9, 1995; revised March 31, 1997. This work was supported by the National Science Council, Taiwan, RO.C., under Contract NSC84-2622-E002-002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Amro El-Jaroudi.

R.-Y. Lyu is with the Department of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan 333, R.O.C. (e-mail: rylyu@ms1.hinet.net, rylyu@cguaplo.cgu.edu.tw).

I.-C. Hong and M.-Y. Lee are with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.

J.-L. Shen is with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.

L.-S. Lee is with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, R.O.C., and the Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.

Publisher Item Identifier S 1063-6676(98)02900-9.

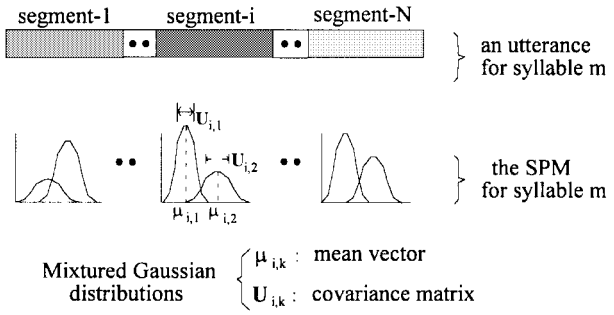

Fig. 1. The concept of SPM.

Below, we shall first present the concept of SPM both in the training and recognition phases. A comparison between "plain" SPM and CHMM (i.e., without special tuning) was then made with extensive experimental results. It will be shown that with the model complexities being the same but without special tuning, SPM performs as well as CHMM for isolated Mandarin base-syllable recognition, while the computation time for SPM is much less than that for CHMM. Different approaches to improve the recognition accuracy and speed for SPM are then proposed. The final results provide a recognition rate of 95.46% (specially tuned CHMM gives 93.89%) at a speed of roughly 45 times faster than CHMM.

II. Segmental Probability Model

In SPM, each utterance of a given base-syllable m is equally divided into N segments (similar to the terminology "state" in CHMM), and the feature vectors in each segment i are modeled by an observation probability distribution function $b_i(\cdot)$, which is composed of several mixtures of Gaussian distributions. Thus the SPM for a given base-syllable m with N segments and M mixtures is denoted as

$$\lambda_m\colon \{b_i(\cdot),\ 1 \le i \le N\}$$

where $b_i(\cdot)$ can have one of the two following forms:

$$b_i(o_t) = \max_{1 \le k \le M} N(o_t;\ \mu_{ik},\ \mathbf{U}_{ik})$$

or

$$b_i(o_t) = \sum_{1 \le k \le M} c_{ik} N(o_t;\ \mu_{ik},\ \mathbf{U}_{ik})$$

where o_t is a feature vector at time t, $N(\cdot;\ \mu_{ik},\ \mathbf{U}_{ik})$ is the multivariate Gaussian distribution with mean vector μ_{ik} and diagonal covariance matrix $\mathbf{U}_{ik}$, and c_{ik} is the mixture weight. Note here i is the segment index, and k is the mixture index. The concept of SPM can be shown as in Fig. 1.

The above SPM model λ_m defines the likelihood of observing a feature vector sequence with length T, denoted as $\mathbf{O} = \mathbf{o}_1\,\mathbf{o}_2\,\cdots\,\mathbf{o}_T$, for a given base-syllable m by

$$P(\mathbf{O}|\lambda_m) = \prod_{1 \le i \le N} \left\{ \prod_{\mathbf{o}_t \in \text{segment } i} b_i(\mathbf{o}_t) \right\}$$

where the feature vector sequence $\mathbf{O}$ is equally divided into N segments. The evaluation process of this likelihood function is shown in Fig. 2.

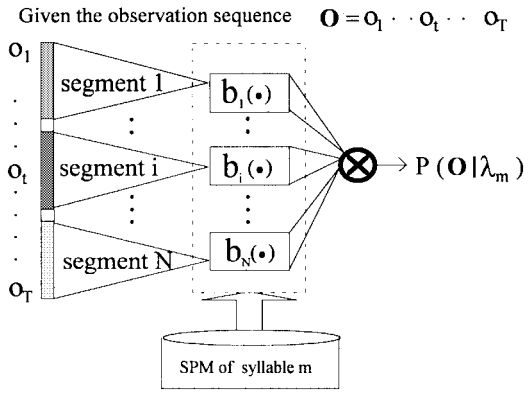

Fig. 2. Evaluation of SPM likelihood function.

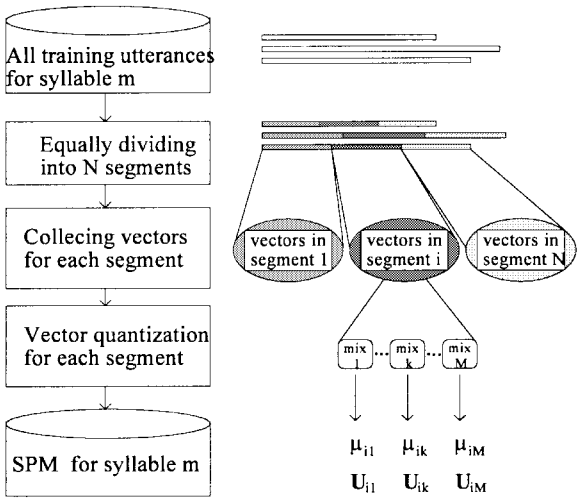

Fig. 3. Training procedure for SPM.

In the recognition phase, after calculating the likelihood functions of observing an unknown feature vector sequence for all possible syllable models, the base-syllable m^* with maximal likelihood was chosen as the recognition output, i.e.,

$$m^* = \underset{1 \le m \le L}{\operatorname{argmax}} P(\mathbf{O}|\lambda_m)$$

where $L = 416$ is the total number of all possible Mandarin base-syllables.

The estimation or training of the parameters for SPM is also simple. All the training utterances for a base-syllable m are first individually divided into N segments of equal length. The feature vectors collected from all the training utterances belonging to the same segment are then clustered into M mixtures using vector quantization technique with Euclidean distance measure. Each of these clusters are then modeled by a Gaussian distribution with the sample mean vector μ_{ik}, and covariance matrix $\mathbf{U}_{ik}$. These sample mean vectors and covariance matrices, $\{\mu_{ik}, \mathbf{U}_{ik}; 1 \le k \le M, 1 \le i \le N\}$, are then used as the SPM parameters for the base-syllable m. The overall training procedure is shown in Fig. 3.

SPM can be understood from an analogy to CHMM. In CHMM, if the state transition probabilities $[a_{ij}]$ are abandoned because of its insignificant effects on the recognition rate, the likelihood of observing a feature vector sequence $\mathbf{O}$ for a model λ is as follows [7]:

$$P(\mathbf{O}|\lambda) = \max_{s} \prod_{t=1}^{T} b_{s_t}(o_t)$$

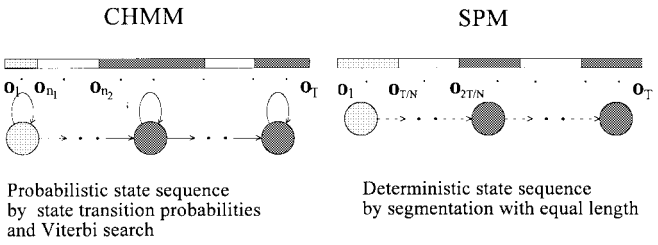

Fig. 4. Difference between SPM and CHMM.

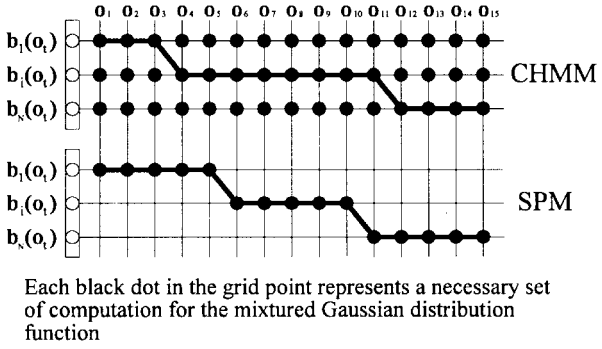

Fig. 5. Comparison of computational complexity between CHMM and SPM in recognition phase.

TABLE I
STATISTICS OF THE SPEECH DATA BASE. EACH SET CONTAINS 1345 ISOLATE MANDARIN TONAL-SYLLABLES (416 BASE-SYLLABLES WITH ALL TONAL VARIATIONS)

		Speaker 1	Speaker 2
Training set	set 1	9.43 min	8.13 min
	set 2	10.29 min	7.66 min
	set 3	9.79 min	8.58 min
Testing set	set 4	10.51 min	8.14 min

where $b_{s_t}(\mathbf{o}_t)$ is the observation probability density function for the feature vector $\mathbf{o}_t$ at state $\mathbf{s}_t$. The maximization is performed over all possible state sequence $\mathbf{S}$, and usually conducted by dynamic programming procedures (like the Viterbi algorithm). In SPM, instead of searching for the optimal state sequence as in CHMM, a particular state sequence was simply specified deterministically by the segments of equal length. From this viewpoint, the SPM likelihood function can be expressed alternatively as

$$P(\mathbf{O}|\lambda) = \prod_{t=1}^{T} b_{s'_t}(o_t)$$

where s'_t is a deterministic segment index at time t, and is defined as $s'_t = \lfloor Nt/T \rfloor$, where $\lfloor x \rfloor$ is the largest integer equal to or smaller than x. The difference between CHMM and SPM can thus be shown as in Fig. 4.

Compared with CHMM, SPM requires only less than $1/N$ (N is the total number of states/segments in CHMM/SPM) of computation in the recognition phase, and only about 1/20 of computation in the training phase as compared to CHMM trained by segmental K-means algorithm [8]. This will be discussed in detail below. First consider the recognition phase. In CHMM, for finding the optimal state sequence, the Gaussian mixture density functions for all states have to be evaluated for each feature vector; but in SPM, since the state sequence is predetermined, it is enough to evaluate the Gaussian

TABLE II
TOP ONE/TOP TEN RECOGNITION RATES FOR THE TWO TESTING SPEAKERS IN DIFFERENT MODEL CONFIGURATIONS WHEN CHMM OR SPM IS USED AS THE RECOGNITION TECHNIQUE, WHERE N IS THE STATE/SEGMENT NUMBER AND M IS THE MIXTURE NUMBER

N \ M	1		2		3	
	CHMM	SPM	CHMM	SPM	CHMM	SPM
3	70.63/99.20	70.03/99.20	80.15/99.85	**80.41/99.85**	82.27/99.85	83.53/100
4	72.27/99.45	72.27/99.45	81.15/99.85	83.27/100	84.31/100	84.91/100
5	75.65/99.45	75.28/99.45	82.90/100	84.09/100	84.31/100	85.28/100
6	77.70/99.50	77.29/99.50	84.24/100	83.57/100	83.98/100	84.68/100
7	80.07/99.85	78.62/99.70	84.68/100	84.20/100	84.37/100	84.91/100

mixture density function only for the determined state at each time, thus only $1/N$ of computation is needed in SPM as compared to CHMM. Furthermore, there is no need to conduct a Viterbi search, which indicates extra computational saving for SPM. So the total computational complexity of SPM is less than $1/N$ of that of CHMM. These discussions are also visualized as in Fig. 5.

In the training phase, for CHMM conventionally adopting the segmental K-means algorithm, the following steps are necessary [8]:

Step 1: model initialization;
Step 2: state sequence segmentation (usually with equal interval);
Step 3: estimate the parameters of the observation probability, via segmental K-means clustering;
Step 4: find the optimal state sequence by Viterbi algorithm;
Step 5: if not converged, go to Step 3.

Usually, it takes about five iterations between Steps 3 and 5 to converge in the above procedure. However, when SPM is used, all that is needed is the first three steps of the above procedure. This reduces the computational load tremendously. In a practical example, assuming $N = 3$, and about 13 training utterances for each base-syllable model, only about 1/20 of computational time as compared to CHMM is enough. This high speed in the training phase is also highly desired, because it makes "on-line" training or learning algorithm practically feasible in a real-time speech recognition system.

III. SPEECH DATA BASE

In this section, the speech data base used in all the following experiments and some front-end signal processing performed on the data base are described. The database was recorded by two male speakers. Each speaker uttered in isolation four sets of the 1345 Mandarin tonal-syllables (the 416 base-syllables plus all possible tonal variations give 1345 tonal-syllables). Therefore, the data base contained 10 760 ($= 2 * 4 * 1345$) isolated syllable templates totally. Some simple statistics of the data base are listed in Table I. For each speaker, three sets of the 1345 tonal-syllables were used as training data and the remaining set is used in testing. Only the base-syllable recognition accuracy (i.e., the tones disregarded) is considered in the tests.

All the recorded materials were obtained in an officelike laboratory environment through a close-talk, noise-canceling AKG C410 microphone. They were digitized with a sampling frequency of 16 kHz via an Ariel ProPort Model 656 analog/digital interface. Each utterance was endpoint-detected first and then preemphasized with a filter $H(z) = 1 - 0.95z^{-1}$. The filtered speech was taken by a 20-ms Hamming window and then cepstral coefficients derived from linear predictive coefficients (LPC-cepstrum) of order 14 were extracted for each 10-ms window shift. Additionally, delta-cepstrum were also computed from every four consecutive frames, i.e., 40 ms.

IV. PERFORMANCE COMPARISON BETWEEN SPM AND CHMM

When the computational complexity of SPM is shown to be much lower than CHMM, a very intuitive guess for the performance comparison is that the accuracy of SPM will be more or less degraded as a natural price paid for the reduced computation. However, extensive experiments presented below will show that this is not true. Here, different model configurations for CHMM and SPM are tested. In these experiments, both cepstral and delta-cepstral coefficients are used. The state/segment number N is changed from 3 to 7, and the mixture number is changed from 1 to 3. The top one and top ten recognition rates for the speech data base mentioned above for the two speakers are listed in Table II.

From Table II it can be seen that the achievable recognition rates for both CHMM and SPM are in fact very close to each other, if the same model configurations are used. These results verify the concept that a deterministic but properly specified state sequence can perform as well as the optimal state sequence found by the Viterbi algorithm. This can be explained by the relatively simple phonetic structure of the target vocabulary of Mandarin base-syllables. Each Mandarin base-syllable is composed of at most three to four phonemes, and the phonetic structure is one of the two types: V and CV, where V represents a vowel or diphthong plus an optional medial and an optional nasal ending, and C represents a consonant. It is thus believed that the equally segmented state sequence is good enough for the discriminative purpose of this vocabulary. On the other hand, the conclusion that both SPM and CHMM give comparable performance is of course limited to the tested vocabulary of Mandarin base-syllables only. It is certainly not necessarily extendible to other vocabulary with more complicated phonetic structures. In Table II, it is also observed that to increase N or M will in general raise the recognition rates in both SPM and CHMM. However, the required computation time and storage space are also proportional to $N \times M$. For this reason, a relatively small N and M are chosen hereafter ($N = 3$, $M = 2$) and further performance improvements are left for additional approaches with less computational burden.

V. ACCURACY IMPROVEMENTS

In this section, three approaches were proposed to achieve accuracy improvements for SPM with little or no computational cost. Considering the confusing subsets mentioned previously, the most discriminative part of the speech signals for Mandarin base-syllable recognition is the very short initial consonants. It is thus easily imagined that to emphasize the importance of such discriminative part in each base-syllable can improve recognition rates. This idea has been successfully used quite early when special training procedures for CHMM was developed for recognizing Mandarin base-syllables [1], [4], [5]. Here, a similar concept has been used in two simple approaches to improve the discrimination among base-syllables. The first is referred to as *nonuniform frame shifting* (NUFS), as

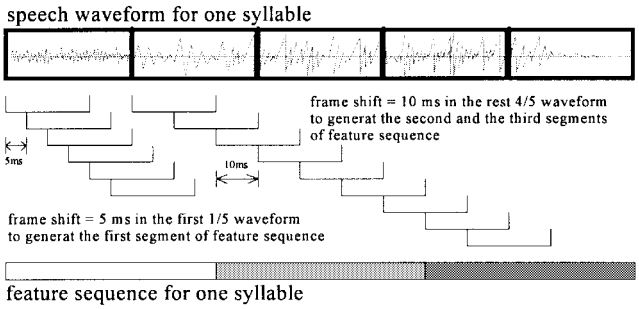

Fig. 6. Nonuniform frame shifting (NUFS).

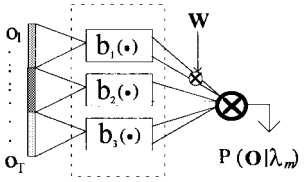

Fig. 7. Weighted likelihood function (WLF).

shown in Fig. 6, which is performed only in the phase of feature extraction. Since the initial consonant of each utterance takes the major responsibility of discrimination, the LPC analysis window can be shifted slower (5 ms versus original 10 ms) in the beginning part (for example, first 1/5 of each base-syllable) than in the remaining part of each utterance. In this way, more feature vectors can be obtained for the important initial consonants. Another approach is referred to as the weighted likelihood function (WLF), i.e., the likelihood values obtained from the beginning part of each utterance (for example, the first segment of feature vectors as shown in Fig. 6) are weighted, as follows:

$$P(\mathbf{O}|\lambda) = \prod_{t=1}^{T} w_t \cdot b_{s_t}(o_t)$$

where $w_t = W$ when $s_t = 1$, and $w_t = 1$ when $s_t > 1$. This is shown in Fig. 7. To determine the optimal weight W, an experiment could be conducted by changing W.

Moreover, the discriminative training based on the *generalized probabilistic descent* (GPD) method could be imposed on the last stage. The basic concept for GPD is to adapt the model parameters such that the likelihood for the correct model increases, and that for the confusing models decreases [9], [10]. In this way, the average error rate will be reduced during the adaptation process. It should be noted that here GPD is applied in the training phase and does not change the recognition phase at all, thus it can be used to improve the recognition rate of the classifier by off-line training at no additional computational cost in recognition.

Three experiments for the above three approaches were conducted and the testing results are listed in Table III and discussed below. Experiment 1 is for the first approach of nonuniform frame shifting (NUFS) (Fig. 6) with $N = 3$, $M = 2$, and the results are listed in the second row of Table III. As compared to the results of plain SPM with $N = 3$, $M = 2$ listed in the first row of the table, it can be seen that the recognition rate has been significantly raised from 80.41 to 89.85%, at the cost of only 1/5 additionally computation time, because the total frame number of each utterance is increased by 1/5. Experiment 2 is for the second approach of WLF when NUFS is already applied, and the results are listed in the next several rows of Table III, where different values of W are tested to find the optimal weighting. The results indicated that the system can achieve

TABLE III
TESTING RESULTS (TOP 1 RECOGNITION RATE AND RELATIVE COMPUTATION TIME) OF THE THREE EXPERIMENTS FOR THE THREE APPROACHES (NUFS, WLF, AND GPD) MENTIONED IN SECTION V TO IMPROVE ACCURACY

		Recognition rate (%)	Relative computation time
Plain SPM (N=3, M=2)		80.41	1
Experiment 1: NUFS	first 1/5: 5ms last 4/5: 10ms	89.85	6/5
Experiment 2: NUFS **plus WLF**	**W=2**	**92.72**	6/5
	W=3	92.31	6/5
	W=4	90.34	6/5
	W=5	88.00	6/5
Experiment 3: NUFS, WLF(W=2) **plus GPD**	10 iterations	94.91	6/5
	20 iterations	**95.80**	6/5
	30 iterations	95.69	6/5
	40 iterations	95.77	6/5

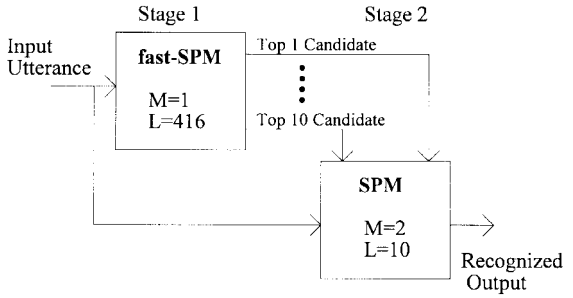

Fig. 8. Two-stage SPM.

a much better recognition rate 92.72% when $W = 2$. These results indicate that the proper use of the phonetic/phonological knowledge of the target vocabulary can significantly improve the performance of a recognizer. Here the emphasis of the beginning segment of each Mandarin syllable using NUFS and WLF can raise the recognition rate from 80.41 to 92.72% with only 20% additional computational cost. However, if we simply try to change the model configuration, i.e., increase N or M as was done in Table II, we can achieve only 85.28% but at 50% additionally computational cost by changing M from 2 to 3, and N from 3 to 7. On the other hand, note that the rate achieved here, 92.72%, is already very close to the best results previously obtained using very finely tuned and specially trained CHMM, 93.89% [5], [6], which was achieved at a much higher computational cost. Experiment 3 is for the third approach of GPD with SPM of $N = 3$, $M = 2$ and the best results obtained previously, i.e., NUFS and WLF ($W = 2$) applied as the initial condition. Different iteration times were executed and tested. The results are listed in the last several rows of Table III. It can be found that the error rate decreased to a local minimum at 20 iteration times, where the recognition rate can be raised to 95.80%. More iteration times do not correct more errors, but are harmless after all. This phenomenon is due to the property of convergence of GPD [11]. Compared with that without GPD adaptation (a recognition rate of 92.72% obtained in Experiment 2 for $W = 2$), the error rate reduction achieved by GPD is 42.32%, i.e., nearly 2/5 of the errors were corrected. This experiment showed that GPD can be well used along with SPM to improve further the performance of SPM effectively.

VI. SPEED IMPROVEMENTS

One of the most important motivations to develop SPM is the speed consideration. In this section, some additional approaches are further developed to speed up the recognition phase of SPM with almost no accuracy degradation, including the fast-SPM algorithm and a two-stage recognition scheme.

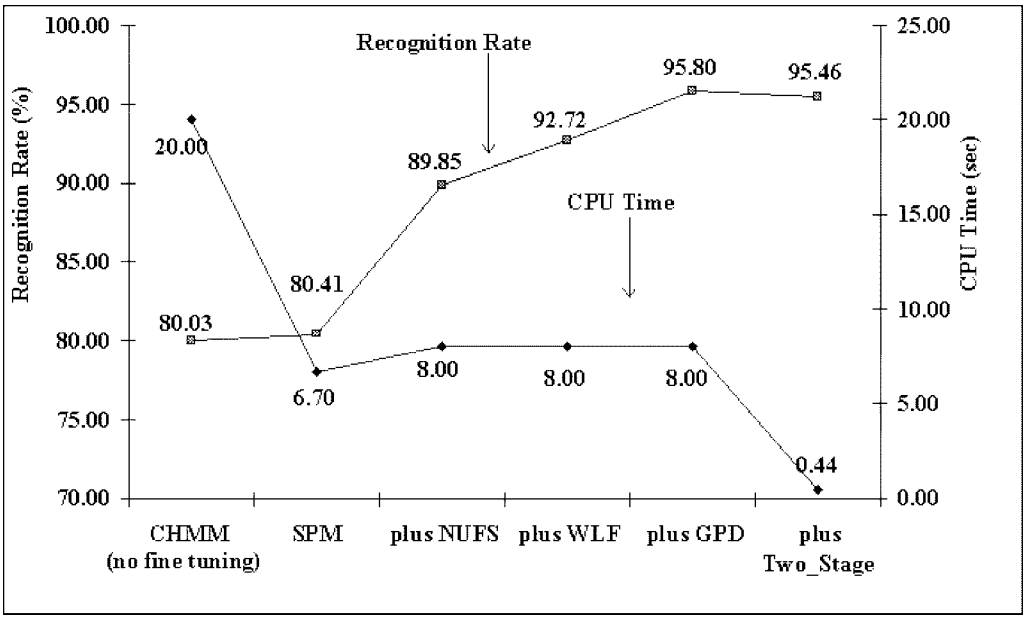

Fig. 9. Recognition rates and the average time needed to recognize one syllable in the Sun Sparc II workstation for the various approaches mentioned in this paper.

TABLE IV
TOP ONE RECOGNITION RATE AND RELATIVE COMPUTATION TIME OF THE ONE-STAGE SCHEME VERSUS TWO-STAGE SCHEME MENTIONED IN SECTION VI

		Recognition rate (%)	Relative computation time
One-stage	SPM (N=3, M=2, L=416) plus NUFS, WLF(W=2) GPD (20 iterations)	95.80	1
Experiment 4: Two-stage	stage1: fast- SPM (N=3, M=1, L=416) stage2: SPM (N=3, M=2, L=10) plus NUFS, WLF(W=2) GPD (20 iterations)	95.46	1/18

TABLE V
OVERALL RESULTS IN MANDARIN BASE-SYLLABLE RECOGNITION

	Recognition rate %	CPU time in Sparc II station (to recognize one syllable)
CHMM (N=3,M=2)	80.03	20.00 sec
SPM(N=3,M=2)	80.41	6.70 sec
plus NUFS	89.85	8.00 sec
plus WLF(W=2)	92.72	8.00 sec
plus GPD(20 iterations)	95.80	8.00 sec
plus two-stage scheme with fast-SPM	95.46	0.44 sec

It was mentioned previously that SPM requires less than $1/N$ computation time compared with CHMM. It will be shown here that if $M = 1$, i.e., there is only one mixture for each segment, a much faster algorithm, referred to as *fast-SPM*, exists due to the equally segmented state sequence of SPM. To describe the fast-SPM, first consider the SPM likelihood function with $M = 1$ as follows:

$$P(\mathbf{O}|\lambda_m) = \prod_{1\le i\le N}\left\{\prod_{o_t\in \text{segment } i} N(\mathbf{o}_t;\,\mu_i,\mathbf{U}_i)\right\}.$$

Note that since there is only one mixture, the mixture index k was dropped here in μ_i and $\mathbf{U}_i$. After some derivations as detailed in the Appendix, the log likelihood function can be expressed as two parts:

$$\log P(\mathbf{O}|\lambda_m) = A(\lambda_m) + B(\mathbf{O}|\lambda_m)$$

where $A(\lambda_m)$ depends only on the model parameters and thus can be evaluated and stored in advance; only $B(\mathbf{O}|\lambda_m)$ depends on both model parameters and the input feature sequence, and thus must be evaluated on-line. Further expansion of $B(\mathbf{O}|\lambda_m)$ shows that it is composed of T (T is the length of the input feature vector sequence $\mathbf{O}$ for a syllable) terms of Gaussian–Euclidean distance functions defined as

$$b(\mathbf{o}_t;\,\mu_i,\,\mathbf{U}_i) = \sum_{d=1}^{D}\frac{(o_{td}-\mu_{id})^2}{\sigma_{id}^2}, \qquad 1\le t\le T$$

where $\mathbf{o}_t = [o_{t1}\cdots o_{td}\cdots o_{tD}]$ is the input feature vector at frame index t with d being the dimension index and D the total number of dimensions in a feature vector $\mathbf{o}_t$, $\mu_i = [\mu_{i1}\cdots\mu_{id}\cdots\mu_{iD}]$ and $\mathbf{U}_i = [\text{diag}\,\sigma_{i1}^2\cdots,\sigma_{id}^2\cdots\sigma_{iD}^2]$ are the mean vector and diagonal covariance matrix for the ith segment of a SPM model. In other words, the computational load of SPM log likelihood function involves T terms of the Gaussian–Euclidean distance function $\mathbf{b}(\mathbf{o}_t;\,\mu_i,\,\mathbf{U}_i)$.

On the other hand, considering the equal-length segmentation of SPM, the individual component and its square in each dimension of

TABLE VI
COMPARISON OF COMPUTATIONAL COMPLEXITY BETWEEN THE PLAIN SPM AND FAST-SPM, $L = 416$, $N = 3$, $D = 14$, AND $T = 53$ USED HERE

	additions	substractions	multiplications	divisions	totals
plain SPM	L(DT-1)	LDT	LDT	LDT	4LDT-L
	308,256	308,672	308,672	308,672	1,234,272
fast-SPM	LDN-L+2DT-2DN	LDN	LDN+DT	LDN	4LDN+3DT-2DN-L
	18,456	17,472	18,214	17,472	71,614
plain SPM / fast − SPM	16.70	17.67	16.95	17.67	17.24

the feature vectors $\mathbf{o}_t$ in a segment can be summed up first, i.e.,

$$O_{id}^{(1)} = \sum_{o_t \in \text{segment } i} \mathbf{o}_{td}$$

and

$$O_{id}^{(2)} = \sum_{o_t \in \text{segment } i} o_{td}^2$$

and then the part $B(\mathbf{O}|\lambda_m)$ discussed above can in fact be further simplified, also detailed in the Appendix, into two parts:

$$\begin{aligned}\log\ P(\mathbf{O}|\lambda_m) &= A(\lambda_m) + B(\mathbf{O}|\lambda_m) \\ &= A(\lambda_m) + C(\lambda_m) + D(\mathbf{O}|\lambda_m).\end{aligned}$$

In this case, $D(\mathbf{O}|\lambda_m)$, the only part that bears the on-line computational load, is composed of only N (N is the number of segments in a SPM model) terms of some kind of "pseudo-Gaussian–Euclidean distance function" $d(i;\ \mu_i,\ \mathbf{U}_i)$ defined as

$$d(i;\ \mu_i,\ \mathbf{U}_i) = \sum_{d=1}^{D} \frac{O_{id}^{(2)} - 2\mu_{id}O_{id}^{(1)}}{\sigma_{id}^2}, \qquad 1 \le i \le N.$$

In other words, the on-line computational load of SPM log likelihood function can involve only N terms of the pseudo-Gaussian–Euclidean distance function $\mathbf{d}(\mathbf{i};\ \mu_i,\ \mathbf{U}_i)$, instead of the T terms of Gaussian–Euclidean distance function $\mathbf{b}(\mathbf{o}_t;\ \mu_i,\ \mathbf{U}_i)$ mentioned above. Since $N \ll T$, this reformulation from frame-by-frame computation of $\mathbf{B}(\mathbf{O}|\lambda_m)$ into segment-by-segment computation of $\mathbf{D}(\mathbf{O}|\lambda_m)$ thus leads to a great reduction of computation on the order of T/N times and is referred to as the fast-SPM. A more detail comparison of the computational complexity between the fast-SPM and the plain SPM is also given in the Appendix.

The above fast-SPM algorithm can reduce the computational time tremendously, but it works only when $M = 1$, in which case SPM can achieve only relatively low top-1 recognition rate (about 70.03% in Table II). Therefore the fast-SPM algorithm is not suitable for direct use. However, although the top-1 recognition rate for $M = 1$ is not high, it is observed from Table II that the top-10 recognition rate for $M = 1$ is very close to 100%. This observation leads to a two-stage recognition scheme as sketched in Fig. 8, in which a fast-SPM algorithm with $M = 1$ is used in the first stage to choose the top ten candidates, then the slower, but more accurate SPM is used in the second stage but performed on only the top ten candidates selected from the first stage. In this way, both stages take very short time in the recognition process, because the first stage employs the fast-SPM algorithm with $M = 1$, while the second stage has to compare the likelihood scores for only ten candidates instead of 416. This two-stage SPM is therefore very fast.

The results of Experiment 4 for the two-stage SPM using the best scheme previously obtained, i.e., $N = 3$, $M = 2$, NUFS, WLF ($W = 2$), GPD (20 iterations), are listed in Table IV. Compared with the original one-stage recognition scheme, the two-stage SPM takes only about 1/18 of computation time, but the recognition rate (95.46%) is only very slightly degraded as compared to the previously mentioned 1-stage scheme (95.80%).

VII. CONCLUSIONS

The overall results of various techniques developed in this paper using SPM are summarized in Table V and also plotted in Fig. 9. The plain SPM, which is a simplified version of CHMM, can do as well as the plain CHMM, around 80.41% accuracy but at less than 1/3 computational load. When we further adopt NUFS and WLF in SPM by considering the characteristics of the target vocabulary at slightly increased computation, the recognition rate was raised to 92.72%, which is actually comparable with the best results ever obtained by other works, using very finely tuned, specially trained CHMM also considering the characteristics of the target vocabulary [4]–[6]. With the discriminative training by GPD method imposed on SPM, the recognition rate can be further improved to 95.80% without changing the recognition speed. Finally, with a two-stage recognition scheme that adopted the fast-SPM algorithm in the first stage, the computation time can be reduced tremendously to about 1/45 of the original CHMM, or an average of 0.44 s per syllable (i.e., almost real time) for recognizing a syllable on a Sun Sparc II workstation, with almost identical accuracy, i.e., 95.46%.

APPENDIX

The Fast-SPM Algorithm: The reformulation of the log likelihood function of SPM is presented here, which leads to the fast-SPM algorithm. For $M = 1$

$$\begin{aligned}\log\ P(\mathbf{O}|\lambda_m) &= \sum_{i=1}^{N} \sum_{t=(i-1)T/N}^{iT/N} \log\ N[\mathbf{o}_i;\ \mu_i^{(m)},\ \mathbf{U}_i^{(m)}] \\ &= -0.5 \sum_{i=1}^{N} \sum_{t=(i-1)T/N}^{iT/N} \sum_{d=1}^{D} \log[2\pi\sigma_{id}^{2(m)}] \\ &\quad - 0.5 \sum_{i=1}^{N} \sum_{t=(i-1)T/N}^{iT/N} \sum_{d=1}^{D} \frac{[O_{td} - \mu_{id}^{(m)}]^2}{\sigma_{id}^{2(m)}} \\ &= A(\lambda_m) + B(\mathbf{O}|\lambda_m)\end{aligned}$$

where $A(\lambda_m)$ and $B(\mathbf{O}|\lambda_m)$ represents, respectively, the two terms in the above equation, m is the model index and $1 \le m \le L$ (=416); λ_m is the parameter set for model m, including the mean and covariance values; D is the dimension of the feature vector. $B(\mathbf{O}|\lambda_m)$ is the part depending on both model parameters and the current utterance. It is easy to see that there are $T * D$ subtractions, $T * D$ multiplications, $T * D$ divisions, and $T * D - 1$ additions to be calculated in $B(\mathbf{O}|\lambda_m)$ for a feature sequence $\mathbf{O}$ with T frames of D-dimensional feature vectors, and the computation of $B(\mathbf{O}|\lambda_m)$ has to be repeated for L times to recognize an unknown utterance. This is the implementation of SPM, which takes roughly $1/N$ of the calculations of conventional CHMM with the same model complexity.

Now, considering the equal-length segmentation of SPM, $B(\mathbf{O}|\lambda_m)$ can be further expanded as follows:

$$\begin{aligned} B(\mathbf{O}|\lambda_m) &= -0.5T\sum_{d=1}^{D}\frac{\mu_{id}^{(m)^2}}{\sigma_{id}^{2(m)}} \\ &\quad - 0.5\sum_{i=1}^{N}\sum_{d=1}^{D}\frac{\left[\sum_{t=(i-1)T/N}^{iT/N} o_{td}^2 - 2\mu_{id}^{m}\sum_{t=(1-1)T/N}^{iT/N} o_{td}\right]}{\sigma_{id}^{2(m)}} \\ &= C(\lambda_m) + D(\mathbf{O}|\lambda_m) \end{aligned}$$

where the on-line computational part $D(\mathbf{O}|\lambda_m)$ can be further expressed as

$$D(\mathbf{O}|\lambda_m) = -0.5\sum_{i=1}^{N}\sum_{d=1}^{D}\frac{[O_{id}^{(2)} - 2\mu_{id}^{(m)}O_{id}^{(1)}]}{\sigma_{id}^{2(m)}}$$

where

$$O_{id}^{(2)} = \sum_{T=(i-1)T/N}^{iT/N} o_{td}^2$$

and

$$O_{id}^{(1)} = \sum_{T=(i-1)T/N}^{iT/N} o_{td}.$$

It is easy to see that there are $N * D$ subtractions, $N * D$ multiplications [$2\mu_{id}^{(m)}$ was calculated and stored in advance], $N * D$ divisions, and $N * D - 1$ additions to be calculated in $D(\mathbf{O}|\lambda_m)$ for a feature sequence $\mathbf{O}$ with T frames of D-dimensional feature vectors, and this computation needs to be repeated by L times for recognizing an unknown utterance. In addition, there are extra T/N multiplications and $(T/N - 1)$ additions for evaluating $O_{id}^{(2)}$, and another $(T/N - 1)$ additions for $O_{id}^{(1)}$, but they depend only on state index i and dimension index d, not on the model index m. This reformulation, referred to as the fast-SPM in this paper, can be interpreted as follows. The plain-SPM is calculating a kernel function (referred to as the *Gaussian–Euclidean distance function* $b(\mathbf{o}_t; \mu_i, \mathbf{U}_i$ in this work) "frame by frame" as the frame index t runs from one to T, but the fast-SPM is calculating a kernel function (referred to as the *pseudo-Gaussian–Euclidean distance function*$d(i; \mu_i, \mathbf{U}_i)$ in this work) "segment by segment" as the segment index i runs from one to N. Since the computational load of the kernel function for both cases are the same, the computational load of the plain-SPM is about T/N times of that of the fast-SPM. The detailed comparison of the computational complexity between the plain-SPM and the fast-SPM are listed in Table VI, with numerical estimations obtained by $L = 416$, $N = 3$, $D = 14$, and $T = 53$. We can see from the table that the fast-SPM takes only about 1/17 of computational time compared to the plain-SPM, which is in turn around 1/45 compared to the conventional CHMM.

REFERENCES

[1] L.-S. Lee *et al.,* "Golden Mandarin (I)—A real-time Mandarin speech dictation machine for chinese language with very large vocabulary," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 158–179, Apr. 1993.

[2] ——, "Golden Mandarin (II)—An improved single-chip real-time Mandarin dictation machine for Chinese language with very large vocabulary," in *Proc. ICASSP'93*, Minneapolis, MN, pp. II.503–II.506.

[3] R.-Y. Lyu *et al.,* "Golden Mandarin (III)—A user-adaptive prosodic-segment-based Mandarin dictation machine for Chinese language with very large vocabulary," in *Proc. ICASSP'95*, Detroit, MI, pp. 57–60.

[4] L.-S. Lee *et al.,* "Special speech recognition approaches for the highly confusing Mandarin syllables based on hidden Markov models," *Comput. Speech Lang.*, vol. 5, pp. 181–201, Apr. 1991.

[5] F.-H. Liu, Y. Lee, and L.-S. Lee, "A direct-concatenation approach to train hidden Markov models to recognize the highly confusing Mandarin syllables with very limited training data," *IEEE Trans. Speech Audio Processing*, vol. 1, Jan. 1993.

[6] Y. Lee and L.-S. Lee, "Continuous hidden Markov models integrating transitional and instantaneous features for Mandarin syllable recognition," *Comput. Speech Lang.*, vol. 7, pp. 247–263, 1993.

[7] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[8] B.-H. Juang and L. R. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1639–1641, Sept. 1990.

[9] W. Chou, B.-H. Juang, and C. H. Lee, "Segmental GPD training of HMM based speech recognizer," in *Proc. ICASSP'92*, San Francisco, CA, pp. I.473–I.476.

[10] P. C. Chang and B.-H. Juang, "Discriminative training of dynamic programming based speech recognizer," in *Proc. ICASSP'92*, San Francisco, CA, pp. I.493–I.496.

[11] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.

Deleted Strategy for MMI-Based HMM Training

Nam Soo Kim and Chong Kwan Un

***Abstract*— In this correspondence, we apply the maximum mutual information (MMI) criterion to discriminative training of hidden Markov model (HMM) parameters. In contrast to the conventional MMI training approach, we adopt the cross-validatory strategy with which the parameters are estimated on a part and assessed on the other parts of the training data. For this purpose, we propose the deleted MMI training method, which performs cross-validatory parameter updating while maintaining the converging behavior of the conventional MMI-based algorithm. The proposed method is compared to the conventional MMI approach in classification of artificial data and in speaker-independent continuous speech recognition, and shows better performance.**

***Index Terms*— Cross validation, hidden Markov model, parameter training.**

I. INTRODUCTION

It is well known that parameter estimation based on a discriminative criterion is more robust to incorrect modeling assumptions than that based on the maximum likelihood (ML) criterion [1]–[8]. Behind this assertion, it is assumed that speech data used for training the model parameters and those used for evaluating recognition performance are generated from the same source. However, in

Manuscript received September 11, 1994; revised August 4, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Amro El-Jaroudi.

N. S. Kim is with the School of Electrical Engineering, Seoul National University, Seoul 151–742, Korea.

C. K. Un is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Taejon, Korea.

Publisher Item Identifier S 1063-6676(98)02918-6.