Computer Speech and Language (1993) **7**, 247–263

Continuous hidden Markov models integrating transitional and instantaneous features for Mandarin syllable recognition

Yumin Lee and Lin-shan Lee

Department of Electrical Engineering, Room 512, National Taiwan University, Taipei, Taiwan, Republic of China

Abstract

Feature parameters describing spectral transitions of speech signals have been properly integrated with the instantaneous features in many different approaches proposed for speech recognition, and significant performance improvements have been attained. Most of these methods are designed for recognition systems based on dynamic time warping (DTW) or discrete hidden Markov models (HMM). However, it has been experimentally shown that for the difficult problem of recognizing the highly confusing Mandarin syllables with limited amount of training data, the performances of DTW and discrete HMM techniques are much worse than that of continuous HMMs. In this paper, the performance of continuous HMMs using one type of transitional features in speaker-dependent recognition of the highly confusing Mandarin syllables is first evaluated and discussed in detail under the constraint of very limited training data. Three approaches are then proposed to integrate the instantaneous and transitional features for recognition systems based on continuous hidden Markov models. They are the most straightforward *concatenation–integration* approach in which the instantaneous and transitional feature vectors are simply concatenated, the *two-maximization* approach in which the output distribution functions for the instantaneous and transitional feature vectors are maximized separately, and the *two-model* approach in which two HMMs respectively for instantaneous and transitional feature vectors are independently trained but the log likelihoods are summed up with proper weighting. After extensive experiments and careful analysis, it is found that the three approaches respectively provide attractive performance under different conditions. For example, with the *two-maximization* approach a recognition rate (93·89%) only slightly lower than the highest achievable rate for the *concatenation–integration* approach (94·36% for $M=5$) can be obtained at a much smaller number of mixtures ($M=2$).

1. Introduction

The recognition of all the 1300 phonologically allowed Mandarin syllables has always been a difficult problem. In fact, Mandarin Chinese is a tonal language, in which every syllable is assigned a tone. It has been found that the vocal tract parameters for the

0885–2308/93/030247 + 17 $08.00/0

syllables are only slightly influenced by the tones (Havie, 1976), and the tones can be separately recognized using pitch contour features. Therefore the differences caused by tones can be disregarded in Mandarin syllable recognition, which reduces the total number of different syllables from 1300 to 408. However, the recognition of all the 408 phonologically allowed Mandarin syllables disregarding the tones is still very difficult. The difficulty is mainly due to the existence of 38 confusing sets in the vocabulary, each of which can have as many as 19 very confusing syllables. A good example is the A-set: {[a]*, [ja], [cha], [sha], [tza], [tsa], [sa], [ga], [ka], [ha], [da], [ta], [na], [la], [ba], [pa], [ma], [fa]}. This is why only speaker dependent recognition is considered currently. Conventionally, each Mandarin syllable can be decomposed into an INITIAL/FINAL format very similar to the consonant/vowel relations in other languages. Here INITIAL means the initial consonant of the syllable, and FINAL means the vowel (probably diphthong) part but including possible medial and nasal ending. It can be found from the above confusing set example that in general all the syllables in a confusing set have a common FINAL but different INITIALs. Table I is a list of all the 408 syllables. The vertical scale of the table lists all the 38 FINALs (including a null FINAL), and the horizontal scale of the table lists all the 22 INITIALs (including a null INITIAL). Therefore every row in the table represents a confusing set, consisting of syllables sharing the same FINAL but with different INITIALs. The A-set mentioned above is listed in the second row of the table. Recognition of these Mandarin syllables becomes even more difficult when only very limited amount of training data is available. However, since it is impractical to require a user new to a speaker-dependent system to produce a large number of training utterances before being able to use the system, very limited amount of training data becomes a necessary constraint. This is why several special recognition approaches for these Mandarin syllables have been developed to efficiently utilize the training data (Liu, Lee & Lee, 1993; Lee *et al.*, in prep.), and very encouraging results have been achieved.

On the other hand, it has been shown that spectral transitions of speech signals play a vital role in speech perception (Furui, 1986). In the past few years, several speech feature parameters that convey transitional spectral information have been proposed (Soong & Rosenberg, 1988). When these transitional features are properly integrated with the instantaneous features in many different approaches proposed (Furui, 1986; Gupta, Lennig & Mermelstein, 1987; Nishimura & Toshioka, 1987; Rabiner, Wilpon & Soong, 1988; Soong & Rosenberg, 1988), significant performance improvements have been attained. Most of these methods are designed for recognition systems based on dynamic time warping (DTW) or discrete hidden Markov models (HMMs). However, it has been experimentally shown that for the difficult problem of recognizing the highly confusing Mandarin syllables with limited training data, the performance of DTW and discrete HMM techniques are much worse than that of continuous HMMs (Liu, Lee & Lee, 1993). Rabiner *et al.* developed a system integrating the instantaneous and transitional speech features in continuous HMMs (Rabiner, Wilpon & Soong, 1988), in which the two types of features are simply concatenated to form augmented feature vectors that contain both the instantaneous and transitional spectral information. As will be shown later in this paper, it is found that such a simple integration approach is not necessarily the best, at least for the problem being discussed here, i.e. the recognition of the very confusing Mandarin syllables with very limited training data, because other attractive approaches can in fact be found in some cases.

* The transliteration symbols used in this paper is Mandarin Phonetic Symbols II (MPS II).

TABLE I. All the 408 Mandarin syllables disregarding the tones. The number indicates the sequence number used in our database

		1	2 j	3 ch	4 sh	5 r	6 tz	7 ts	8 s	9 g	10 k	11 h	12 ji	13 chi	14 shi	15 d	16 t	17 n	18 l	19 b	20 p	21 m	22 f
1			1	2	3	4	5	6	7														
2	a	8	9	10	11		12	13	14	15	16	17				18	19	20	21	22	23	24	25
3	o	26																		314	315	316	317
4	e	27	28	29	30	31	32	33	34	35	36	37				38	39	40	41				
5	ai	42	43	44	45		46	47	48	49	50	51				52	53	54	55	56	57	58	
6	eh	59																					
7	ei	60	61		62		63		64	65		66				67		68	69	70	71	72	73
8	au	74	75	76	77	78	79	80	81	82	83	84				85	86	87	88	89	90	91	
9	ou	92	93	94	95	96	97	98	99	100	101	102				103	104	105	106		107	108	109
10	en	110	111	112	113	114	115	116	117	118	119	120				121	122	123	124	125	126	127	128
11	an	129	130	131	132	133	134	135	136	137	138	139						140		141	142	143	144
12	ang	145	146	147	148	149	150	151	152	153	154	155				156	157	158	159	160	161	162	163
13	eng	164	165	166	167	168	169	170	171	172	173	174				175	176	177	178	179	180	181	182
14	i	183											184	185	186	187	188	189	190	191	192	193	
15	u	194	195	196	197	198	199	200	201	202	203	204				205	206	207	208	209	210	211	212
16	iu	213											214	215	216			217	218				
17	ia	219											220	221	222				223				
18	ie	224											225	226	227	228	229	230	231	232	233	234	
19	iai	235																					
20	iau	236											237	238	239	240	241	242	243	244	245	246	
21	iou	247											248	249	250	251		252	253			254	
22	ian	255											256	257	258	259	260	261	262	263	264	265	
23	in	266											267	268	269			270	271	272	273	274	
24	iang	275											276	277	278			279	280				
25	ing	281											282	283	284	285	286	287	288	289	290	291	
26	ua	292	293	294	295					296	297	298											
27	uo	299	300	301	302	303	304	305	306	307	308	309				310	311	312	313				
28	uai	318	319	320	321					322	323	324											
29	uei	325	326	327	328	329	330	331	332	333	334	335				336	337						
30	uan	338	339	340	341	342	343	344	345	346	347	348				349	350	351	352				
31	uen	353	354	355	356	357	358	359	360	361	362	363				364	365		366				
32	uang	367	368	369	370					371	372	373											
33	ueng	374	375	376	377	378	379	380	381	382	383	384				385	386	387	388				
34	iue	389											390	391	392			393	394				
35	iuan	395											396	397	398				399				
36	iun	400											401	402	403								
37	iung	404											405	406	407								
38	er	408																					

In this paper, the performance of continuous HMMs using one type of transitional feature parameter in speaker-dependent recognition of the highly confusing Mandarin syllables is first evaluated and discussed in detail under the constraint of very limited training data. Three different approaches are then proposed to integrate the instantaneous and transitional features for recognition systems based on continuous hidden Markov models. They are the most straightforward *concatenation–integration* approach in which the instantaneous and transitional feature vectors are simply concatenated, the *two-maximization* approach in which the output distribution functions for the instantaneous and transitional feature vectors are maximized separately, and the *two-model* approach in which two HMMs respectively for instantaneous and transitional feature vectors are independently trained but the log likelihoods are summed up with proper weighting. Extensive experiments are performed to test and compare the performance of these proposed approaches for the problem discussed here, and the results are carefully analysed. The conditions under which each approach will be the most attractive are also discussed. For example, with the *two-maximization* approach a recognition rate (93·89%) only slightly lower than the highest achievable rate for the *concatenation–integration* approach (94·36% for $M=5$) can be obtained at a much smaller number of mixtures ($M=2$). In the following, the speech database and the hidden Markov model with the training approach to be used will be presented in Sections 2 and 3, respectively. The effectiveness of the transitional spectral features used is then assessed in Section 4. Three different methods of integrating these features with the instantaneous features and the experimental results are then discussed in Sections 5 and 6, and finally a conclusion is given in Section 7.

2. The speech database

In this section the speech database and some initial processing performed on the speech data will be presented first. They will be used in all the experiments to be discussed in the following. The database contains two collections of data of two male speakers. Each includes six utterances for each of the 408 Mandarin syllables. Therefore the database contains a total of $4896=408\times6\times2$ syllable templates. All syllables are uttered in isolation, in the high-level tone (usually referred to as the first tone in Mandarin) and in random sequences. The times for recording the speech data were distributed in five days.

All the recorded materials are obtained in an office-like laboratory environment without special sound-proof treatment. They are low-pass filtered and digitized through a MASSCOMP-5400 workstation, and then stored in the hard disk for further processing. The sampling frequency is 10 kHz. After end-point detection (Rabiner & Sambur, 1975) is performed for each syllable, 20-ms Hamming window is applied every 7 ms to obtain the autocorrelation coefficients with a preemphasis factor of 0·95. LPC-based cepstral analysis is then performed on the autocorrelation coefficients to extract the first ten cepstral coefficients which are used as the instantaneous features. Regression analysis is further applied to obtain transitional features. The linear regression coefficients are evaluated using the following equation (Furui, 1986):

$$r_t^m = \frac{\sum_{l=-3}^{3} c_{(t+l)}^m \cdot l}{\sum_{l=-3}^{3} l^2}$$

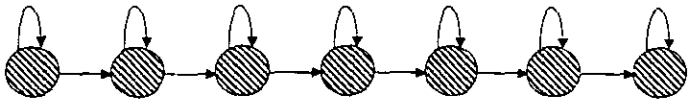

Figure 1. The state transition topology of the HMMs used.

where c_t^m is the m-th cepstral coefficient at time t, and r_t^m is the linear regression coefficient for the m-th cepstral coefficient at time t. The training data for each speaker consist of five utterances for each of the 408 syllables and the remaining one utterance is used in testing. In other words, the testing data for each speaker includes one utterance for each of the 408 syllables, and the recognition rates discussed in the following are the average for the two speakers. Note that here only five training utterances are available for each syllable, which makes the training task very difficult, although it already takes a very long time for a new speaker to produce $2040 = 408 \times 5$ training utterances.

On the other hand, 408 training utterances for the 408 syllables need to be segmented into INITIAL parts and FINAL parts for use in further processing. Such segmentation is in fact very difficult, especially when the INITIAL is some unaspirated plosives such as /b/, /d/, /g/, or some nasals or liquids such as /m/, /n/, /r/, /l/, etc. It was found that the spectral transition measure previously proposed (Svendsen & Soong, 1987) was a useful feature to define the boundary between INITIAL and FINAL parts even for syllables starting with those difficult INITIALs mentioned above for which the average magnitude difference function (AMDF) (Ross *et al.*, 1974) is not useful at all. The spectral transition measure is defined below:

$$C(t) = \sum_{m=1}^{10} \left(\sum_{l=-3}^{3} w_l \cdot c_{(t+l)}^m \cdot l \right)^2 \tag{1}$$

where $C(t)$ is the spectral transition measure at time t and w_l is a window with length of 7 frames.

3. The hidden Markov model with partitioned Gaussian mixtures

The hidden Markov models used in this paper are left-to-right continuous mixture HMMs with 7 states, as shown in Fig. 1. The output probability density function (pdf) of a state j is the *partitioned Gaussian mixtures* (PGM) function:

$$b_j(\mathbf{o}_t) = \frac{1}{M} \max_{m=1,2,\ldots,M} \{b_{jm}(\mathbf{o}_t)\} \tag{2}$$

where $\mathbf{o}_t$ is the feature vector, M is the total number of mixtures, and $b_{jm}(\cdot)$ is the m-th mixture pdf of state j, which is assumed to be a multi-dimensional Gaussian distribution. This type of pdf is similar in concept to, and is in fact motivated by, the *partitioned Gaussian autoregressive mixtures* (PGAM) density function previously proposed (Juang & Rabiner, 1985). In this type of mixture density, the feature vector space can be considered to be implicitly partitioned into clusters. Each cluster is defined by a Gaussian pdf. As shown in Equation (2), the cluster to which a feature vector $\mathbf{o}_t$ belongs

is found by a nearest-neighbour criterion (Juang & Rabiner, 1985). This in fact resembles the vector quantization (VQ) operation. Such a VQ analogy will be used in later discussions in this paper. The parameters to be re-estimated are the transition probabilities a_{ij}, mixture mean vectors $\boldsymbol{\mu}_{jm}$, and mixture covariance matrices $\mathbf{R}_{jm}$. The formula for maximum likelihood re-estimation of these parameters using the Baum–Welch training algorithm can be easily obtained by following the previous derivations (Juang, 1985). In particular, let the training utterances for one syllable be $\mathbf{O}^{(v)}=\{\mathbf{o}_1^{(v)},\mathbf{o}_2^{(v)},\ldots,\mathbf{o}_{T(v)}^{(v)}\}, v=1,2,\ldots,V$, where $\mathbf{o}_t^{(v)}$ is the feature vector of the v-th utterance at time t, V is the total number of utterances, and $T^{(v)}$ is the length of the v-th training utterance. Also, let $s_t^{(v)}=i$ denote the event that the v-th utterance is in state i at time t. The re-estimated parameters $\tilde{a}_{ij}$, $\tilde{\boldsymbol{\mu}}_{jm}$, $\tilde{\mathbf{R}}_{jm}$ can then be obtained as follows:

$$\tilde{a}_{ij}=\frac{\sum_{v=1}^{V}\sum_{t=1}^{T(v)-1} f(\mathbf{O}^{(v)},s_t^{(v)}=i,s_{(t+1)}^{(v)}=j|\Lambda)}{\sum_{v=1}^{V}\sum_{t=1}^{T(v)-1} f(\mathbf{O}^{(v)},s_t^{(v)}=i|\Lambda)} \tag{3}$$

$$\tilde{\boldsymbol{\mu}}_{jm}=\frac{\sum_{v=1}^{V}\sum_{t=1}^{T(v)}[f(\mathbf{O}^{(v)},s_t^{(v)}=j|\Lambda)\cdot\mathbf{o}_t^{(v)}\cdot q_{jm}(\mathbf{o}_t^{(v)})]}{\sum_{v=1}^{V}\sum_{t=1}^{T(v)}[f(\mathbf{O}^{(v)},s_t^{(v)}=j|\Lambda)\cdot q_{jm}(\mathbf{o}_t^{(v)})]} \tag{4}$$

$$\tilde{\mathbf{R}}_{jm}=\frac{\sum_{v=1}^{V}\sum_{t=1}^{T(v)}[f(\mathbf{O}^{(v)},s_t^{(v)}=j|\Lambda)\cdot(\mathbf{o}_t^{(v)}-\boldsymbol{\mu}_{jm})\cdot(\mathbf{o}_t^{(v)}-\boldsymbol{\mu}_{jm})^T\cdot q_{jm}(\mathbf{o}_t^{(v)})]}{\sum_{v=1}^{V}\sum_{t=1}^{T(v)}[f(\mathbf{O}^{(v)},s_t^{(v)}=j|\Lambda)\cdot q_{jm}(\mathbf{o}_t^{(v)})]} \tag{5}$$

In these equations, $f(\cdot)$ is used to denote the density function, and Λ denotes the HMM. For example, the quantity $f(\mathbf{O}^{(v)},s_t^{(v)}=i,s_{(t+1)}^{(v)}=j|\Lambda)$ is the probability density of observing $\mathbf{O}^{(v)}$ with a transition from state i at time t to state j at time $t+1$, given the current model parameter set Λ. The quantity $f(\mathbf{O}^{(v)},s_t^{(v)}=i|\Lambda)$ is the probability density of observing $\mathbf{O}^{(v)}$ with the process being in state i at time t. Equation (3) has an intuitive interpretation of simply being the fractional count of transitions from state i to state j averaged over all sequences. $q_{rm}(\cdot)$ in Equations (4) and (5) is the nearest neighbourhood function defined as follows:

$$q_{jm}(\mathbf{o}_t)=\begin{cases}1 \text{ if } \operatorname{argmax}_{l=1,2,\ldots,M}\{b_{jl}(\mathbf{o}_t)\}=m\\ 0 \text{ otherwise}\end{cases}$$

Since the amount of training data is limited, the covariance matrices are assumed to be diagonal. From Equation (4), it is clear that the re-estimated mean vector of the m-th mixture of state j is the weighted centroid of the vectors assigned to that particular mixture. The covariance matrix in Equation (5) have a similar interpretation. It is worth noting that the re-estimation algorithm for the mean vectors is similar to the LBG algorithm [Linde, Buzo & Gray, 1980) used to train VQ codebooks.

Since we are dealing with the recognition of highly confusing Mandarin syllables using very limited amount of training data, special training approaches are required. Several special approaches have been proposed to utilize the training data efficiently, and very good results have been obtained with the *revised three-pass* training approach (Lee *et al.*, in prep.) using HMMs with PGAM output density. In this approach, as shown in the block diagram in Fig. 2, one set of training utterances are segmented into INITIAL and FINAL parts. These segmented utterances are used to train the INITIAL and FINAL

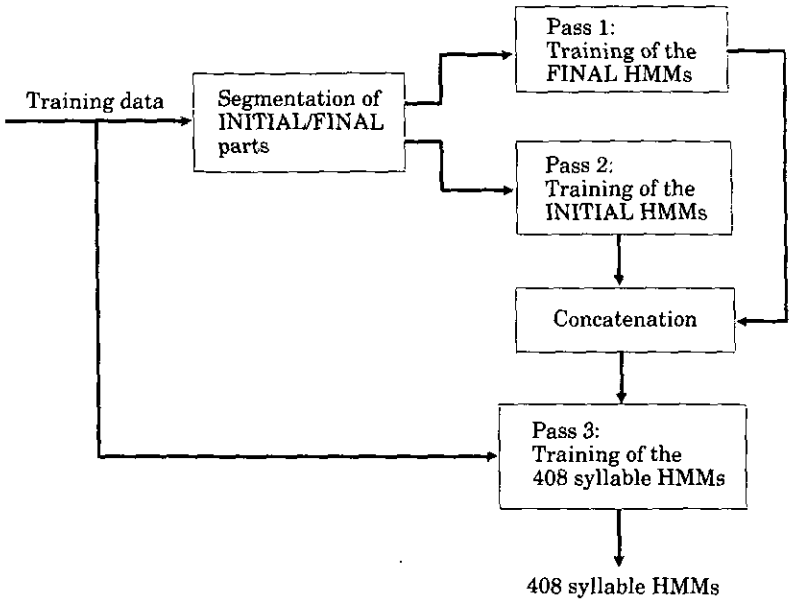

Figure 2. Block diagram of the *revised three pass* training algorithm.

HMMs in the first two passes. They are then concatenated to form 408 syllable HMMs. In the third pass, these 408 syllable HMMs are taken as the initial values, and all the segmented as well as the unsegmented training utterances are used in the Baum–Welch iterations to refine the model parameters, with the parameters for the INITIAL and FINAL parts of the syllable HMMs eventually re-estimated separately. This method is adopted to train the HMMs in the experiments to be discussed in this paper.

A series of preliminary experiments were performed to evaluate the performance of the PGM models and of the HMMs with conventional Gaussian mixtures (GM) pdf's (Rabiner, Juang, Levinson & Sondhi, 1985) using cepstral coefficients as features:

$$b_j(\mathbf{o}_t) = \sum_{m=1}^{M} c_{jm} b_{jm}(\mathbf{o}_t) \tag{6}$$

where M, $b_{jm}(\cdot)$ and $\mathbf{o}_t$ are the same as in Equation (2), and c_{jm} are non-negative real numbers satisfying the condition

$$\sum_{m=1}^{M} c_{jm} = 1$$

The results of these preliminary experiments are listed in Table II. From these results one can see that the PGM models can always achieve better performance than the GM models. This can be attributed to the following reasons. First, the number of model parameters to be re-estimated is significantly smaller for the PGM models than that for the GM models. More specifically, for the PGM models there is no need to re-estimate the parameters $c_{jm}, j = 1,2,\ldots,N, m = 1,2,\ldots,M$. Therefore for the typical case of $N = 7, M = 5$, the number of parameters of the PGM models is that for the GM models

TABLE II. Recognition rates of HMMs using different output pdfs

		Top n rates (%)				
M	model	1	2	3	4	5
2	GM	84·56	92·16	94·36	95·10	95·10
	PGM	85·54	94·61	96·32	97·30	97·30
3	GM	81·62	91·91	94·12	94·61	95·34
	PGM	86·52	94·36	95·83	96·57	96·85
4	GM	79·90	92·40	93·87	94·12	94·61
	PGM	81·86	92·65	95·10	96·57	97·30
5	GM	82·60	92·38	93·63	94·61	95·10
	PGM	84·07	93·14	95·10	95·34	95·59
6	GM	81·13	89·22	91·91	93·87	94·85
	PGM	82·11	93·87	96·08	96·32	97·06

minus 35. Thus for the problem here with only very limited training data, the PGM models with fewer parameters will of course outperform the GM models with more parameters. Secondly, it seems that the PGM model is more suitable for Mandarin syllable recognition than the GM models because of its partitioned-type of output pdf (Lee *et al.*, in prep.). One reason for this is that as shown in Equation (6) the observation density for GM models is the weighted average of the mixture components, while for the PGM models in Equation (2) only the mixture component producing the maximum likelihood will be picked up as the observation density. For highly confusing Mandarin syllables, the smoothing effect of the weighted average operation tends to obscure the subtle differences among confusing syllables and make the model less discriminating. Another reason is that, as mentioned earlier, modelling with this partitioned-type of mixture density resembles the VQ operation, therefore an HMM with the partitioned-type of output pdf in fact resembles the operation of finite state vector quantization (FSVQ) (Juang & Rabiner, 1985), which had been shown to be extremely effective for the recognition of confusing Mandarin syllables (Lee *et al.*, in prep.). Therefore although the GM models have been found equally successful in many speech recognition tasks, the performance could be quite different if applied under special conditions.

4. The effectiveness of regression coefficients

In order to test the effectiveness of the regression coefficients as a feature for recognition, a series of experiments were conducted in which the regression coefficients were used alone, while the number of mixtures M is varied from 2 to 9. To provide a baseline for comparison, the cepstral coefficients were also used alone in the recognition experiments. The results are listed in Table III. The top-1 recognition rates as functions of M are also plotted in Fig. 3(a).

Two observations are worth noting in Table III. First, when cepstral coefficients are used alone the top-1 recognition rates degrade in general as M is increased beyond 5. This indicates that the discriminating ability of these models becomes worse when too many mixtures are used. The reason for this is twofold. On the one hand, the amount of training data is so scarce that when M is increased beyond 5, the estimated model parameters are no longer accurate. On the other hand, Mandarin syllables are highly

TABLE III. Recognition rates of HMMs using cepstral coefficients (CEPS) and regression coefficients (REGRS) alone

		Top n rates (%)				
M	feature	1	2	3	4	5
2	CEPS	85·54	94·61	96·32	97·30	97·30
	REGRS	87·25	95·10	96·08	97·30	98·28
3	CEPS	86·52	94·36	95·83	96·57	96·81
	REGRS	86·76	94·61	95·83	97·06	97·55
4	CEPS	81·86	92·65	95·10	96·57	97·30
	REGRS	84·56	95·10	95·34	96·81	97·30
5	CEPS	84·07	93·14	95·10	95·34	95·59
	REGRS	86·52	94·61	95·10	95·59	97·06
6	CEPS	82·11	93·87	96·08	96·32	97·06
	REGRS	84·31	94·36	95·34	95·83	96·81
7	CEPS	80·15	92·16	94·12	95·59	97·79
	REGRS	86·27	94·85	95·34	96·32	97·06
8	CEPA	80·15	92·16	94·12	94·85	96·32
	REGRS	84·07	94·36	96·57	98·04	99·02
9	CEPS	80·39	91·67	94·12	95·10	95·83
	REGRS	84·07	94·36	96·57	98·04	99·02

confusing, and thus very elaborate models are necessary to discriminate the subtle differences between confusing syllables.

Secondly, using regression coefficients alone yields significantly higher recognition rates than using cepstral coefficients alone, with the exception that $M=3$ yields almost equal top-1 rates. To illustrate this more clearly, typical example results of $M=5$ are plotted in Fig. 3(b). This phenomenon is different from the results reported by Furui (1986), which stated that regression coefficients are only slightly more efficient than the instantaneous cepstral coefficients, if sufficient training data are available. Although the recognition approach adopted by Furui (1986) is entirely different from the approach used here, a possible reason for such difference is that here we are recognizing syllables of a very confusing vocabulary using very limited amount of training data. Under these circumstances the robustness of estimated model parameters (e.g. the means and covariances) becomes very important, and it seems that the models trained using the regression coefficients have in fact more robust model parameters with respect to limited training data. By robustness here we mean that reasonably accurate estimations of the means and covariances and so on in the models are obtainable even if only very limited training data are available. Consequently, the model parameters obtained here using regression coefficients are more accurate than those using cepstral coefficients. This supposition is further supported by Fig. 3(c) which shows the improvements in top-1 recognition rates using the regression coefficients with respect to using the cepstral coefficients. From Fig. 3(c), one can see that the increase in top-1 recognition rates is in general larger for large M. In other words, as M increases, more parameters are to be re-estimated, and when the amount of training data is insufficient, the need for robustness in model parameters becomes more crucial.

To probe further, we can also examine the output pdfs of the HMMs. Typical output distributions for the HMMs using cepstral and regression coefficients as features are

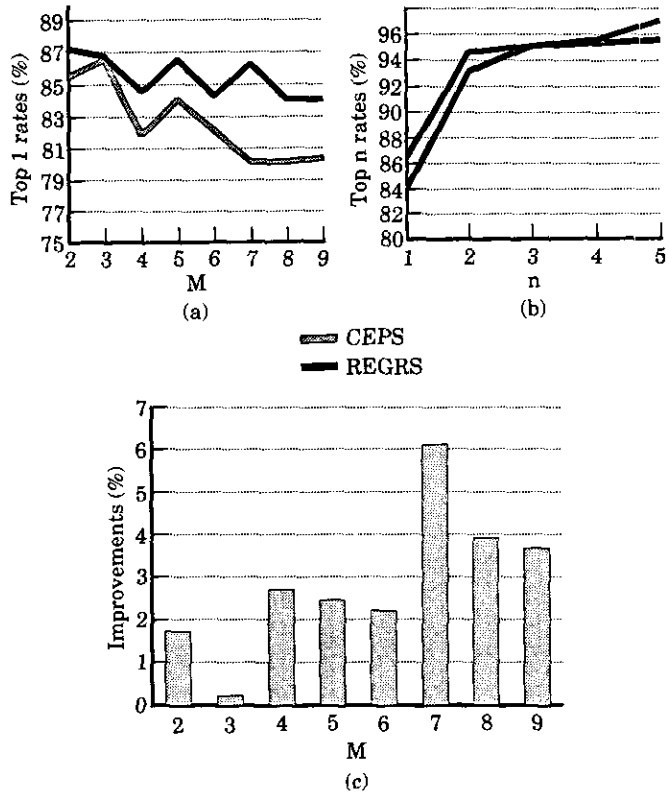

Figure 3. Recognition results of HMMs using cepstral coefficients (CEPS) and regression coefficients (REGRS) alone. (a) Top-1 rates for M = 2 up to 9; (b) Top n rates for M = 5; (c) Improvements attained by using the regression coefficients.

plotted in Fig. 4(a) and (b). From Fig. 4 one can see that the output distributions for models trained by the regression coefficients are typically much more compact than those by the cepstral coefficients. In other words, the intraspeaker variations in the regression coefficients are very small. This accounts for the robustness of the models using regression coefficients with respect to limited training data, because when the same number of samples is used, the estimation of the statistical parameters tends to be more accurate for distributions with smaller variances than for distributions with larger variances. Therefore since intraspeaker variations in regression coefficients are small, reasonably accurate estimations of the model parameters can still be obtained regardless of the scarcity of the training data.

5. Integration of transitional and instantaneous features

It has been shown that the instantaneous and transitional spectral features are largely uncorrelated (Gupta, Lennig & Mermelstein, 1987; Soong & Rosenberg, 1988). Therefore, the information conveyed in these two types of features are basically complementary, and thus can be used jointly to improve recognition performance. Several approaches to appropriately integrate these two types of features have been proposed

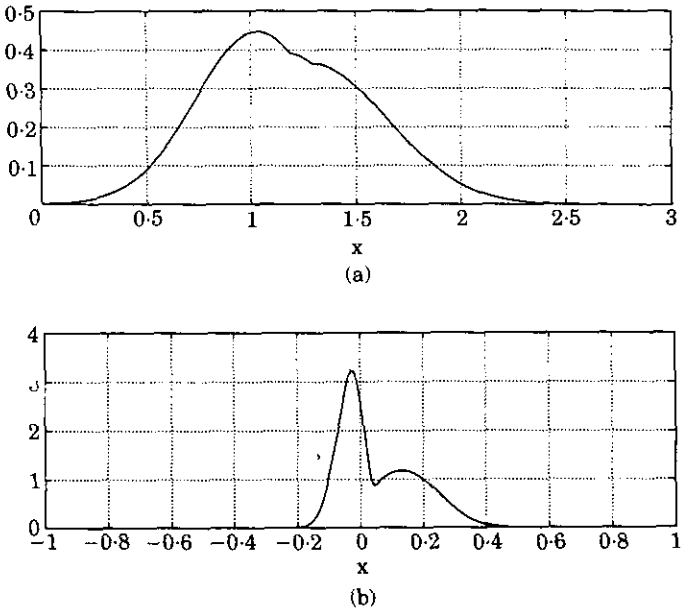

Figure 4. Typical output distributions of HMMs using (a) cepstral and (b) regression coefficients as features.

(Furui, 1986; Gupta *et al.*, 1987; Nishimura & Toshioka, 1987; Soong & Rosenberg, 1988). These approaches were designed for VQ-based or for discrete-HMM-based recognition systems. In this section, different approaches to integrate the regression and cepstral coefficients for continuous-HMM-based recognition systems will be presented and discussed.

At least three different approaches of incorporating the cepstral and regression coefficients are possible. The most straightforward approach is to directly concatenate the P dimensional vectors of the cepstral and regression coefficients into a $2P$-dimensional feature vector, where P is the order of the cepstral analysis. The state output pdfs of the HMMs follow Equation (2), where $\mathbf{o}_t$ now refers to the concatenated feature vectors and $b_{jm}(\cdot)$ are $2P$-dimensional Gaussian distributions. This will be referred to as the *concatenation–integration* approach.

The second approach is to modify Equation (2) so that vectors of cepstral coefficients and vectors of regression coefficients are accounted for separately. Specifically, let $\mathbf{c}_t$ and $\mathbf{r}_t$ denote the cepstral and regression coefficient vectors, respectively. The state output pdf now becomes

$$b_j(\mathbf{c}_t,\mathbf{r}_t)=\frac{1}{M\times L}\underset{m=1,2,\ldots,M}{max}\{b^c_{jm}(\mathbf{c}_t)\}\max_{l=1,2,\ldots,L}\{b^r_{jl}(\mathbf{r}_t)\} \tag{7}$$

In Equation (7), $b^c_{jm}(\cdot)$ and $b^r_{jl}(\cdot)$ are P-dimensional Gaussian distributions of the cepstral and regression coefficient vectors, respectively, and M and L the number of mixtures of the cepstral coefficient and regression coefficient distributions, respectively. The maximum likelihood re-estimation formula for the transition probabilities follows Equation (3), where $\mathbf{o}_t$ is simply the concatenation of $\mathbf{c}_t$ and $\mathbf{r}_t$. The mean vectors $\tilde{\boldsymbol{\mu}}^c_{jm}$ and $\tilde{\boldsymbol{\mu}}^r_{jl}$ and the covariance matrices $\tilde{\mathbf{R}}^c_{jm}$ and $\tilde{\mathbf{R}}^r_{jl}$, where superscripts c and r denote the cepstral and

regression coefficient distributions, respectively, can be re-estimated using Equations (4) and (5) with $\mathbf{o}_t$ substituted with the corresponding feature vector. Here, the covariance matrices for the distributions $b_{jm}^c(\cdot)$ and $b_{jl}^r(\cdot)$ are again assumed to be diagonal because of the insufficiency of the training utterances. Note that in this approach, although one single HMM is used, the parameters of the cepstral and regression coefficient vector distributions are re-estimated separately. This approach will be referred to as the *two-maximization* approach.

The last approach is to train independently two HMMs for each syllable. The first HMM uses the cepstral coefficients as features, while the second uses the regression coefficients as features. During recognition the log likelihood of the test utterance is the weighted sum of the log likelihoods obtained from these two HMMs using the corresponding features. Quantitatively, let $\mathbf{O}^c$ and $\mathbf{O}^r$ denote the time sequences of cepstral and regression coefficient vectors extracted from the test utterance, respectively. The log likelihood of this utterance given the syllable HMMs Λ^c,Λ^r is defined as

$$L(\mathbf{O}^c,\mathbf{O}^r|\Lambda^c,\Lambda^r) \stackrel{\text{def}}{=} w\cdot L(\mathbf{O}^c|\Lambda^c)+(1-w)\cdot L(\mathbf{O}^r|\Lambda^r) \tag{8}$$

where $L(\cdot|\Lambda^c)$ and $L(\cdot|\Lambda^r)$ are log likelihoods given the syllable HMM Λ^c with M mixtures using cepstral coefficients as features and the syllable HMM Λ^r with L mixtures using regression coefficients as features, respectively. $0\leqslant w\leqslant 1$ is a real number used to specify the relative weight between the two HMMs. This method is similar to the *word-level integration* previously developed (Gupta *et al.*, 1987), with the difference that continuous, instead of discrete HMMs, are used here, and that log likelihoods are combined with weightings. This approach will be referred to as the *two-model* approach. In general here the HMMs Λ^c and Λ^r can be completely different. This means that they can be independently optimized in terms of various parameters such as number of states, state transition topology, type of state output pdf, and the number of mixtures.

It was mentioned in Section 3 that the PGM densities resemble the vector quantization operation in that the feature vector space is implicitly partitioned into clusters. From this point of view, using PGM models to represent the feature vectors obtained from the first *concatenation–integration* approach mentioned above is conceptually similar to jointly quantizing the cepstral and regression coefficient vector spaces, i.e., quantizing the cepstral and regression coefficient vector spaces using a single joint codebook. On the other hand, integrating the cepstral and regression coefficient vectors using the *two-maximization* or *two-model* approaches resembles, in some sense, the separate vector quantization (SPVQ) (Nakamura & Sjikano, 1989), which is a type of product code VQ. For these two latter methods, the cepstral and regression coefficient vector spaces are implicitly partitioned separately into clusters using two different codebooks. In the *two-maximization* approach, one can see from Equation (7) that the effective number of mixtures is $M\times L$, while the number of parameters to be re-estimated is proportional to $M+L$, where M and L have been defined below Equation (7). The same idea applies to the *two-model* approach. This means that models with large number of mixtures are *effectively* obtained while only a small number of parameters are to be re-estimated. This is particularly advantageous when the available amount of training data is limited, and will be referred to as the *separate quantization effect* in the following discussions. Practically, however, the effective number of mixtures may not be as large as it seems for the following reasons. First, although the cepstral and regression coefficients are largely uncorrelated, correlation between these two types of features still more or less exist.

TABLE IV. Recognition rates using the concatenation–integration (CI), two-maximization (2MAX) and two-model (2MOD) approaches

M	Integration Approach	Top n rates (%)				
		1	2	3	4	5
	CI	92·65	97·30	98·04	98·04	98·53
2	2MAX	93·87	98·04	98·53	98·53	98·53
	2MOD	91·67	97·79	98·53	98·77	99·02
	CI	93·14	98·04	98·77	99·02	99·02
3	2MAX	93·14	98·28	98·77	99·02	99·26
	2MOD	90·44	96·57	98·53	99·02	99·02
	CI	87·01	97·55	99·02	99·51	99·75
4	2MAX	91·91	97·79	98·53	98·53	98·77
	2MOD	89·95	96·32	97·30	97·79	98·04
	CI	94·36	99·02	99·26	99·51	99·75
5	2MAX	90·93	97·06	98·04	98·28	99·02
	2MOD	89·46	96·08	97·30	97·55	97·79
	CI	92·16	97·30	97·79	98·28	98·77
6	2MAX	91·42	96·81	97·30	97·30	98·04
	2MOD	89·95	96·57	98·04	98·28	98·28
	CI	93·87	97·79	98·04	98·28	98·53
7	2MAX	91·42	97·06	97·55	97·79	97·79
	2MOD	89·95	96·08	97·30	97·79	98·28
	CI	92·16	97·55	97·79	98·28	98·53
8	2MAX	87·01	97·55	99·02	99·51	99·75
	2MOD	91·67	97·55	98·28	98·53	99·02
	CI	92·16	97·55	97·79	98·53	98·53
9	2MAX	91·18	97·55	97·55	98·04	98·04
	2MOD	92·16	98·04	99·02	99·51	99·51

Secondly, although two maximization operations are independently performed in the *two-maximization* approach as indicated in Equation (7), a single HMM is used, hence the implicit quantization of the two vector spaces are not really independent for this approach. Lastly, the instantaneous and dynamic features actually convey phonetic information through mutual interaction (Furui, 1986). In other words, although using the *two-maximization* and *two-model* approaches improves the efficiency in training data utilization, the models thus obtained may be inaccurate. It therefore remains to be verified experimentally in the next section whether the advantages really dominate in the present problem.

6. Experimental results and discussions

Extensive simulation experiments were performed to evaluate the effectiveness of the proposed integration approaches in the recognition of the highly confusing Mandarin syllables under the constraint of limited amount of training data. For the *two-maximization* and *two-model* approaches, PGM models with $M = L$ are chosen for simplicity. The weight w for the *two-model* approach is optimized to 0·4 by a series of preliminary experiments.

The recognition rates using the proposed integration approaches with M varying from 2 to 9 are listed in Table IV. Comparing these results with those listed in Table III, one

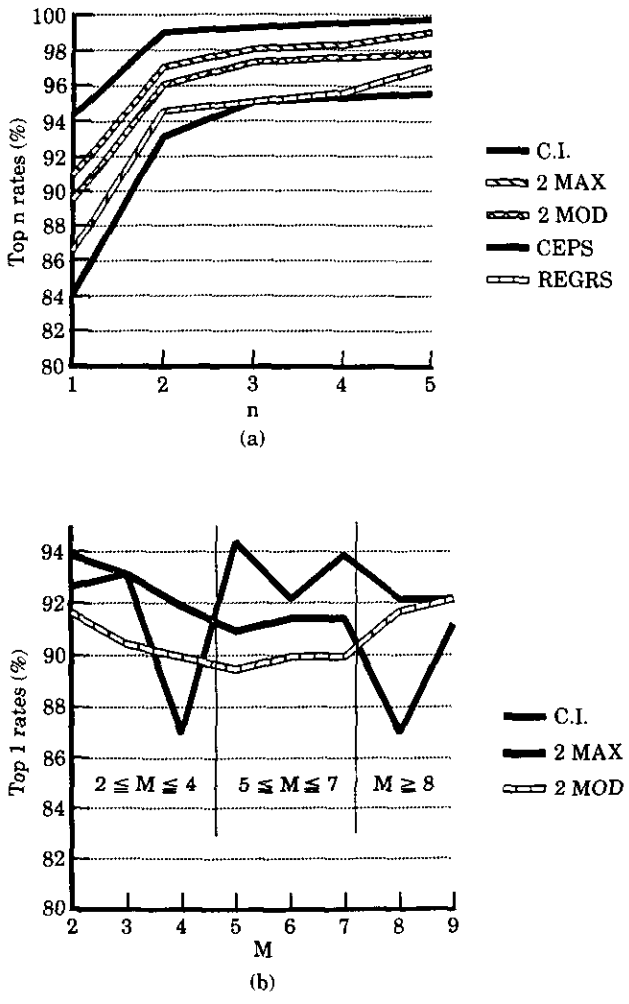

Figure 5. Recognition results using the concatenation–integration (CI), two-maximization (2MAX) and two-model (2MOD) approaches. (a) Top n rates for M = 5; (b) Top-1 rates for M = 2 up to 9.

can easily appreciate the significant improvement brought about by incorporating the regression and the cepstral coefficients together. The top-1 rates now exceed 90% in most of the cases, which are 7–10% higher than the results obtained using the regression coefficients alone, and 8–13% higher than the results achieved using the cepstral coefficients alone. The top-5 rates are now on the order of 98–99% and are 2–3% and 1–2% respectively higher than the results obtained using the cepstral and the regression coefficients alone. The highest top-1,2,3,4, and 5 recognition rates are all achieved by using the *concatenation–integration* approach with M set to 5. The recognition results for this particular case together with the results obtained by using the cepstral and regression coefficients alone are plotted in Fig. 5(a) for illustration purposes.

Although the *concatenation–integration* approach seems to be the best from Fig. 5(a), this is not always true in fact. For example, as can be found in Table IV, the *two-*

maximization approach can achieve a top-1 rate of 93·87% for $M=2$ only, which is only about 0·5% lower than the highest top-1 rate here, 94·36% achieved by the *concatenation–integration* approach with $M=5$. Furthermore, when the top-1 recognition rates as a function of M are plotted in Fig. 5(b) for the different approaches, some interesting phenomena can be observed. First, for $2 \leqslant M \leqslant 4$, the *two-maximization* approach gives a better performance than the *concatenation–integration* and *two-model* approaches. Also, the top-1 recognition rates achieved by the *two-model* approach are in general 2–3% lower than those achieved by the *concatenation–integration* and *two-maximization* approaches. The only exception to this fact is the case of $M=4$ in which the top-1 recognition rate of the *two-model* approach is almost 3% higher than that of the *concatenation–integration* approach. Secondly, for $5 \leqslant M \leqslant 7$, the top-1 recognition rates achieved by the *two-model* approach are still 2–3% lower than those achieved by the other approaches. The situation, however, is reversed for the *concatenation–integration* and *two-maximization* approaches: now the top-1 recognition rates of the *concatenation–integration* approach are higher than those of the *two-maximization* approach. Lastly, for $M \geqslant 8$, the top-1 recognition rates achieved by the *two-model* and *concatenation–integration* approaches become comparable, and are higher than those achieved by the *two-maximization* approach. A closer comparison between the *concatenation–integration* and *two-model* approaches in the last six rows of Table IV reveals that for $M=8$ and $M=9$ the top-1 rates of the *two-model* approach are respectively 0·5% lower than and equal to that of the *concatenation–integration* approach. However, one also can see that for both cases the top-2,3,4 and 5 rates of the *two-model* approach are almost all higher than that of the *concatenation–integration* approach. Another interesting observation is that for both the *concatenation–integration* and *two-maximization* approaches the top-1 rates for $M \geqslant 8$ is lower than those for $5 \leqslant M \leqslant 7$. However, on the contrary, the top-1 rates of the *two-model* approach are higher for $M \geqslant 8$ than for $5 \leqslant M \leqslant 7$. Based on all these observations, one can conclude that for small number of mixtures, i.e., $2 \leqslant M \leqslant 4$, the *two-maximization* approach in general yields the highest top-1 recognition rates. For medium number of mixtures ($5 \leqslant M \leqslant 7$), the *concatenation–integration* is the most attractive. For M larger than or equal to 8, the *two-model* and *concatenation–integration* approaches both give very good results, with the *two-model* approach performing slightly better in terms of the top-2,3,4 and 5 rates.

These conclusions can be interpreted from the implicit quantization point of view as discussed previously in Section 5. Because of the high degree of confusion in the Mandarin syllables, HMMs with small number of mixtures are incapable of discriminating the subtle differences among the confusing syllables. This explains why for $M \leqslant 4$ the *concatenation–integration* approach cannot yield very good performance. For the *two-maximization* and the *two-model* approaches, however, the *effective* number of mixtures is relatively larger due to the separate quantization effect discussed in Section 5. Hence both the *two-maximization* and *two-model* approaches have the potential to outperform the *concatenation–integration* approach. However, as can be found in Fig. 5(b), the *two-model* approach in fact does not outperform the *concatenation–integration* approach. On the contrary, its recognition rate is 2 to 3% lower than the *concatenation–integration* approach. This can be attributed to the fact that for the *two-model* approach, two independent HMMs are used. As was mentioned in Section 5, the instantaneous and transitional features convey phonetic information through mutual interaction. Using two independent HMMs for the cepstral and regression coefficients respectively, as in the *two-model* approach, in fact does not make any use of the mutual interaction.

Apparently although some advantages are obtainable using the *two-model* approach, the disadvantages actually dominate here. Of course the same problem also exists in the *two-maximization* approach because in this approach two maximization operations are independently performed as indicated in Equation (7). However, the problem is to a lesser extent for this approach because only one single HMM, instead of two independent HMMs, is used. This is why for the case of $2 \leqslant M \leqslant 4$ the *two-maximization* approach performs the best, for which the separate quantization effect really dominates.

For $5 \leqslant M \leqslant 7$, the models obtained using the *concatenation–integration* approach possess relatively good discriminating capability because the number of mixtures is large enough now, thus are capable of distinguishing the subtle differences among the highly confusing syllables. The separate quantization effect, of course, still exists in the *two-maximization* and *two-model* approaches. However, for the *two-maximization* approach, using large M does not necessarily imply that the effective number of mixtures is very large. This is because the implicit quantization of the cepstral and regression coefficient vector spaces are not really independent, as mentioned previously in Section 5. Increasing M may only lead to very little increase in the effective number of mixtures. On the other hand, because of the limited amount of training data, a large M may lead to inaccurate estimation of model parameters. Therefore in the case of $5 \leqslant M \leqslant 7$, the losses brought about by the insufficiency of training data in fact surpass the gains introduced by the separate quantization effect for this approach. For the *two-model* approach, using large M actually results in a large effective number of mixtures because two independent HMMs are used. However, the problems of the mutual interactions between the cepstral and regression coefficients and limited training data still dominate, hence the performance of the *two-model* approach is only satisfactory for the cases of $5 \leqslant M \leqslant 7$.

For the cases of large number of mixtures, i.e., $M \geqslant 8$, the amount of training data is so insufficient that for the *concatenation–integration* approach, increasing M gains nothing. It is clear that models obtained using this approach are inaccurate and is incapable of discriminating the very confusing syllables. The same argument applies to the *two-maximization* approach. It was stated earlier and can be found in Fig. 5(b) that the top-1 recognition rates achieved by the *concatenation–integration* and *two-maximization* approaches for $M \geqslant 8$ are in fact lower than those for $5 \leqslant M \leqslant 7$. However, the top-1 recognition rates achieved by the *two-model* approach for $M \geqslant 8$ become significantly higher than those for $5 \leqslant M \leqslant 7$. This indicates that now the separate quantization effect actually dominates in this approach. In other words, because of the separate quantization effect, models with good discriminating capability are obtainable despite the fact that only very limited amount of training data are available. This is equivalent to saying that when $M \geqslant 8$ the *two-model* approach provides a means of very efficiently utilizing the very limited amount of training data.

7. Conclusion

The performance of the previously proposed regression coefficients (Furui, 1986) as features for the recognition of the highly confusing Mandarin syllables is evaluated under the constraint of very limited training data. It is experimentally found that the regression coefficients are very efficient under these circumstances. Three different approaches are then proposed to integrate the regression coefficients with the cepstral coefficients. These approaches are suitable for continuous-HMM-based recognition systems. Significant improvements in recognition rates are achieved, and each of the

three approaches is shown to provide very good performance under different conditions. After careful analysis and discussions, it is found that in general the *two-maximization* and *concatenation–integration* approaches are very attractive for small (less than 4) and medium (from 5 to 7) numbers of mixtures, respectively. For number of mixtures $M \geqslant 8$, on the other hand, the *two-model* and *concatenation–integration* approaches can both achieve very good performance, which is much better than that of the *two-maximization* approach. In fact, in our experiments the *two-model* approach performs slightly better than the *concatenation–integration* approach in terms of the top-2,3,4 and 5 rates. In any case, there does exist a full spectrum of design options in choosing the approaches and parameters to optimize the desired performance and computation complexity. For example, with the *two-maximization* approach a recognition rate (93·89%) only slightly lower than the highest achievable rate for the *concatenation–integration* approach (94·36% for $M = 5$) can be obtained at a much smaller number of mixtures ($M = 2$).

References

Havie, J. M. (1976). *Acoustical Studies of Mandarin Vowels and Tones*, Cambridge, Cambridge University Press.

Liu, F.-H., Lee, Y. & Lee, L.-S. (1993). A direct-concatenation approach to train hidden Markov models to recognize the highly confusing Mandarin syllables with very limited training data. To appear on *IEEE Transactions on Speech and Audio.*

Lee, L.-S., *et al.*, Special speech recognition approaches for the highly confusing Mandarin syllables based on hidden Markov models, *Computer Speech and Language* (accepted).

Furui, S. (1986). On the role of spectral transition for speech perception. *Journal of the Acoustic Society of America*, **80(4)**, 1016–1025.

Soong, F. K. & Rosenberg, A. E. (1988). On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **36(6)**, 871–879.

Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **34(1)**, 52–59.

Gupta, V. N., Lennig, M. & Mermelstein, P. (1987). Integration of acoustic information in a large vocabulary word recognizer. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1987, 697–700.

Nishimura, M. & Toshioka, K. (1987). HMM-based speech recognition using multi-dimensional multi-labeling. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1163–1167.

Rabiner, L. R., Wilpon, J. G. & Soong, F. K. (1988). High performance connected digit recognition using hidden Markov models. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1988, 119–122.

Rabiner, L. R. & Sambur, M. R. (1975). An algorithm for determining the endpoints of isolated utterance. *AT&T B.S.T.J.*, **54(2)**.

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

Ross, M. J., *et al.* (1974). Average magnitude difference function pitch extractor. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **22(10)**, 353–362.

Juang, B.-H. & Rabiner, L. R. (1985). Mixture autoregressive hidden Markov models for speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **33(6)**, 1404–1413.

Juang, B.-H. (1985). Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains. *AT&T Technical Journal*, **64(6)**, 1235–1249.

Linde, Y., Buzo, A. & Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications*, **28(1)**, 84–94.

Rabiner, L. R., Juang, B.-H., Levinson, S. E. & Sondhi, M. M. (1985). Recognition of isolated digits using hidden Markov models with continuous mixture densities. *AT&T Technical Journal*, **64(6)**, 1211–1234.

Nakamura, S. & Shikano, K. (1989). Speaker adaptation applied to HMM and neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 89–92.