Phonemic Hidden Markov Models with Continuous Mixture Output Densities for Large Vocabulary Word Recognition

L. Deng, P. Kenny, M. Lennig, V. Gupta, F. Seitz, and P. Mermelstein

Abstract—**In this study we demonstrate the effectiveness of phonemic hidden Markov models with Gaussian mixture output densities (mixture HMM's) for speaker-dependent large-vocabulary word recognition. Speech recognition experiments show that for almost any reasonable amount of training data, recognizers using mixture HMM's consistently outperform those employing unimodal Gaussian HMM's. With a sufficiently large training set (e.g., more then 2500 words), use of HMM's with 25-component mixture distributions typically reduces recognition errors by about 40%. We also found that the mixture HMM's outperform a set of unimodal generalized triphone models having the same number of parameters. Previous attempts to employ mixture HMM's for speech recognition proved discouraging because of the high complexity and computational cost in implementing the Baum–Welch training algorithm. We show how mixture HMM's can be implemented very simply in the unimodal transition-based frameworks by the device of allowing multiple transitions from one state to another.**

I. Introduction

For HMM-based large vocabulary speech recognition, use of phonemes as the basic unit of speech is attractive for several significant reasons. First and foremost, since there are only about 40 phonemes in English, phonemic HMM's can be trained adequately with practically feasible amounts of speech data. Second, phonemically based recognizers enable the user to add new words to the recognition vocabulary with great convenience. Third, the phonemic HMM approach allows the recognizer to focus on those regions of speech which are inherently confusable. This avoids potential masking of discrimination by random variation in the phonemically common portion of the words [1]. Finally, use of phonemic HMM's facilitates development of heuristically based fast search algorithms to reduce the computational requirements in decoding the spoken words [2].

We have evaluated the performance of an 86 000-word speaker dependent isolated word recognizer where each phoneme is represented by one HMM with unimodal Gaussian output densities. De-

Manuscript received April 23, 1989; revised June 26, 1990.

L. Deng was with INRS-Telecommunications, Montreal, Que., Canada H3E 1H6. He is now with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada.

P. Kenny, M. Lennig, V. Gupta, F. Seitz, and P. Mermelstein are with INRS-Telecommunications, Montreal, Que., Canada H3E 1H6.

IEEE Log Number 9144717.

spite the above advantages, the recognition rate achievable by our system was seriously limited no matter how much training data was used. The major problem considered responsible for the poor performance is the excessively high variability in acoustic realization of a phoneme attributable principally to the phonetic contexts in which it occurs. Much higher variances are observable in the Gaussian distribution associated with the HMM states near the phoneme boundaries than with those near the center [3].

A stochastic model for a phoneme must be sufficiently flexible to capture the great variability in its acoustic realization. If a phoneme is always encountered in the same context, it can be argued that a unimodal Gaussian HMM provides an adequate description. To allow for contextual variation, a more general model appears desirable. One way of accommodating the highly variable phoneme is to relax the unimodal assumption for the HMM output densities. In this correspondence, we report our experiments with phonemic HMM's having Gaussian mixture output densities (abbreviated as mixture HMM's hereafter); for previous applications of mixture HMM's to speech recognition see [4]–[7]. While phonemes represented by mixture HMM's are still assumed independent of their phonetic contexts, the multimodal output densities in mixture HMM's can be expected to provide a more complete representation of the observed acoustic parameter variation in the HMM state than those in unimodal Gaussian HMM's. In this correspondence, we address the problem of how to make efficient use of free parameters in HMM-based speech recognition, and provide experimental evaluation of the mixture HMM's in an 86 000-word recognizer.

II. Mixture HMM's

The HMM we use to represent a phoneme is based on an underlying left-to-right Markov chain having 4 to 10 states. The parameters that characterize the HMM are as follows:

1) $A = [a_{ij}]$, $i, j = 1, 2, \cdots, N$, the state transition matrix of the Markov chain, where a_{ij} is the transition probability from state i to state j and N is the number of states. For the left-to-right HMM we have used, we assume $a_{ij} = 0$, for $j < i$ and $j > i + 2$.

2) A probability density on the observation vectors which is defined for each transition in the Markov chain. For each transition, the output distribution is a Gaussian mixture having a density of the form

$$b_{ij}(\boldsymbol{O}) = \sum_{m=1}^{M} c_{ij}^{(m)} g_{ij}^{(m)}(\boldsymbol{O}) \tag{1}$$

where $\boldsymbol{O}$ is the observation vector, $c_{ij}^{(m)}$ is the weight for the mth mixture component associated with a transition from state i to state j. $g_{ij}^{(m)}(\boldsymbol{O}) = N[\boldsymbol{O}, \boldsymbol{\Theta}_{ij}^{(m)}, \Sigma]$ is a Gaussian density for the mth mixture component associated with a transition from state i to j. In $g_{ij}^{(m)}(\boldsymbol{O})$, the mean vector $\boldsymbol{\Theta}_{ij}^{(m)}$ is distinct for each state transition and for each mixture component. The covariance matrix Σ, specific to each phoneme, is assumed to be common to all mixture components and to all state transitions, in order to save computation and to make its estimate more reliable.

The training algorithm for mixture HMM's can be found in [8]. However, we found it more convenient to adopt a slightly different formulation of the model which enables us to train it by means of the well-known unimodal Gaussian training algorithm, with minor modifications to existing software. The idea is that for each mixture HMM, it is possible to construct an equivalent unimodal HMM by allowing multiple parallel transitions between pairs of states, as in Fig. 1. Associated with each transition in the mixture HMM there is a transition probability p, and, for each $m = 1, \cdots, M$, a mixture weight $c_{ij}^{(m)}$ and a Gaussian density; in the equivalent unimodal model, this transition is replaced by a set of M parallel branches whose transition probabilities are $pc_{ij}^{(m)}$ and whose output distributions are the corresponding mixture components. We will refer to this unimodal as a multiple-branch HMM.

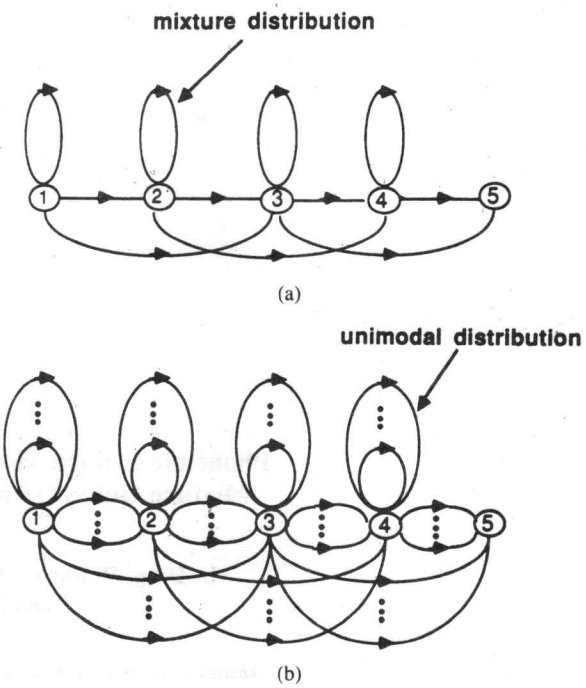

Fig. 1. (a) State diagram of a mixture HMM. The distribution associated with each transition is a mixture of M Gaussian distributions. (b) State diagram of an equivalent multiple-branch HMM. Each transition in (a) is replaced by M parallel branches. The distribution associated with each branch is unimodal Gaussian.

III. Speech Recognition Experiments

This section describes a series of experiments to evaluate the effectiveness of the phonemic mixture HMM's discussed in the previous sections. The task is speaker-dependent isolated-word recognition on an 86 000-word English vocabulary.

Training and test data from nine speakers, four males and five females, were recorded in a quiet sound booth using a Crown PZM microphone, low-pass filtered at 7.1 kHz and sampled at the rate of 16 kHz. A Hamming window with a width of 25.6 ms is applied at intervals of 10 ms; each frame is represented by a 15-dimensional vector consisting of mel-frequency cepstral coefficients and their differences over time [2]. The data consists of natural language sentences spoken as sequences of isolated words, selected randomly from magazines, books, office correspondence and newspapers. The number of words spoken by each speaker varied from about 2000 to 5000. The partition of these words into training and test sets will be indicated in describing the experiments.

The structure of our large-vocabulary isolated-word recognizer has been described previously [2], [9], [10]. Briefly, the recognition process consists of word-endpoint detection, a fast search algorithm to generate a list of most likely word choices (300 choices on average), the computation of exact likelihoods for these choices, and the use of a uniform language model or a trigram language model trained on 57-million words of text.

A. Recognition Results with 25-Component Mixture HMM's

A list of experimental results from nine speakers is presented in Table I. For each speaker, the number of words used in training

TABLE I
RECOGNITION RESULTS USING 25-COMPONENT MIXTURE HMM'S FOR NINE SPEAKERS

Speaker (sex)	Training Size (Words)	Test Size (Words)	Performance		
			Fast Search	Uniform Language	Trigram Language
AM(M)	2742	565	96.4%	69.5%	86.2%
CA(F)	2343	1090	96.5%	83.1%	91.1%
FS(M)	2322	1014	98.3%	91.6%	95.4%
JM(M)	1664	587	97.0%	75.4%	92.8%
LM(F)	2353	863	96.2%	76.3%	88.6%
MA(F)	1600	586	97.8%	85.7%	95.1%
MG(F)	2338	564	97.7%	86.3%	93.6%
ML(M)	2066	580	98.3%	85.5%	93.8%
NM(F)	1299	967	96.0%	81.6%	90.0%

TABLE II
COMPARISON OF RECOGNITION RATES FOR UNIMODAL PHONEMIC HMM'S, UNIMODAL TRIPHONE HMM'S, AND 25-COMPONENT MIXTURE PHONEMIC HMM'S. UNIFORM AND TRIGRAM LANGUAGE MODELS

Speaker (test size)	Training Size	Unimodal HMM's		Mixture HMM's		Triphone HMM's	
		Uniform	Trigram	Uniform	Trigram	Uniform	Trigram
CA(female)	717	67.9%	80.6%	69.7%	80.5%	69.7%	82.5%
(1090 words)	1532	70.2%	85.0%	81.0%	90.0%	78.1%	88.0%
	2347	70.6%	86.6%	86.1%	94.2%	80.8%	90.4%
	3098	70.8%	86.8%	86.0%	94.0%	82.7%	91.3%
	3880	70.7%	86.7%	85.9%	94.0%	83.0%	91.9%
AM(male)	1100	54.4%	78.0%	55.4%	79.0%	53.4%	77.0%
(698 words)	2039	68.2%	81.0%	73.9%	87.0%	72.0%	83.0%
	2742	68.7%	81.9%	76.7%	89.7%	76.1%	86.4%
MA(female) (586 words)	1600	79.0%	90.4%	86.2%	92.5%	83.3%	89.6%

and test sets, and the recognition accuracy achieved at various stages of the mixture-HMM recognizer are shown in the table as separate columns. The test data comprises 6816 words in total.

Shown in the fourth column is the performance of the fast search strategy based on the syllabic graph search algorithm which has been described elsewhere [2]. The performance is measured by the percentage of test words which are selected as one of 300 candidate words from among 86 000 words in the vocabulary. On average over the nine speakers (weighted by the size of the test sets), the proportion of words missed by the fast search algorithm is 3.0%. This result, obtained by the use of mixture HMM's in the syllabic graph, is significantly better than that by the use of unimodal Gaussian HMM's.

Shown in the fifth and sixth columns are the recognition accuracy with the use of the uniform and the trigram language model, respectively. The accuracy is measured by the percentage of test words correctly identified by the recognizer as the top word choice. Homophone confusions are counted as errors when the trigram language model is used, but not when the uniform language model is used. The weighted average of the recognition rates for the uniform and trigram language models are 82.2% and 91.8%, respectively.

B. Comparison of Recognition Rates Obtained with Unimodal Phonemic HMM's, Mixture Phonemic HMM's, and Unimodal Generalized Triphone HMM's for Various Training Set Sizes

Our work on mixture HMM's grew out of earlier work on generalized triphone modeling which has been described elsewhere [11]. We defined 25 generalized triphones for each phoneme based on a five-way classification of left and right phonetic contexts and trained unimodal Gaussian HMM's for each triphone. We then used the mean vectors from these models as an initialization for training mixture HMM's (one per phoneme). The two sets of HMM's thus have the same number of parameters. It came as something of a surprise to discover that the mixture HMM's outperformed the triphone HMM's in almost every instance both with the uniform and trigram language models, as indicated in Table II (columns 5–8).

Columns 3 and 4 contain recognition rates for unimodal phonemic HMM's. Except for the experiment performed on speaker CA with 717 words of training data, the performance of mixture models is substantially better. Note that for both unimodal and mixture models performance improves as the amount of training data increases but that the improvement saturates much earlier for the unimodal models. Although the number of parameters in the mixture HMM's is much larger than in the unimodal HMM's we have not seen any evidence of under-training problems in the mixture case. (Recall that we are tying the covariance matrix across all mixture components on all transitions in each of the phoneme models so that all of the extra parameters in the mixture HMM's are accounted for by the mean vectors.)

C. Effects of Different Initializations in Training Mixture HMM's

The parameter of HMM's trained by the Baum–Welch algorithm depend on the initialization used. In the case of mixture HMM's there is good reason to be concerned that recognition performance

TABLE III
COMPARISON OF LOG LIKELIHOODS ON TRAINING DATA (AT CONVERGENCE AFTER FIVE ITERATIONS) AND OF RECOGNITION ACCURACY USING MIXTURE HMM'S TRAINED FROM DIFFERENT INITIAL POINTS. UNIFORM LANGUAGE MODEL

Initialization Method	Log Likelihood on Training Data	Percent Correct
Context-dependent HMM's	−1753296	76.3%
Mixture HMM's from another speaker	−1769875	76.7%
Partition tokens	−1759080	76.0%

may be adversely affected by an inappropriate initialization. For instance, if the parameters of two mixture components are set equal at initialization, they will remain equal on all subsequent iterations, effectively reducing the number of mixture components by one.

We have experimented with three different ways of initializing the 25-component mixture HMM's. The first method is to use the mean vectors of the triphone models referred to in the previous section as initial estimates of the modes of the mixture distributions. In the second method we simply use mixture HMM's trained from another speaker. In the third method, frames of the training data are aligned with mixture components in two steps and the estimates of the mean vector of each of the mixture components is obtained by averaging all the frames aligned with it. First, the frames in each token are aligned with transitions in a set of unimodal Gaussian HMM's and we impose the same alignment with the transitions in the set of mixture models to be initialized. Second, frames are assigned to the mixture components associated with the transitions by randomly associating a number between 1 and 25 with each token.

Likelihoods on training data and the corresponding recognition rates with the uniform language model using the mixture HMM's obtained by the three different initializations are compared in Table III for a male speaker (AM). The training and test sizes are 2742 and 698 words, respectively. For the second method of initialization, the prototype mixture HMM's are from a female speaker (CA), trained with 3880 words. The results shown in Table III suggest that there is not much difference in the HMM's obtained by the three methods. In fact the recognition errors in each of the three cases turned out to be much the same.

D. Effects of Varying the Number of Mixture Components

Table IV shows the result of a series of recognition experiments carried out to determine the effect of varying the number of mixture components when the size of the training set is fixed.

The experiments were carried out for one male speaker (AM) with a training set of 2742 words and a test set of 698 words and for one female speaker (CA) with training and test sets of 2342 and 1090 words, respectively. The recognition rates quoted were obtained using the uniform language model.

Observe that increasing the number of components from 25 to 38 leads to no improvement in the case of either speaker. It may be that an improvement could be obtained with more data; another possibility is that with 25 components per mixture we may have already exhausted the possibilities of phonemic mixture HMM's.

IV. CONCLUSION

In this correspondence we have shown how mixture HMM's can be implemented very simply in the unimodal transition-based framework by the device of allowing multiple parallel transitions between pairs of states. With this formulation, mixture HMM's can be trained by a direct application of the Baum–Welch algorithm for unimodal Gaussian HMM's. The mixture HMM's described in this correspondence are formally different from the tied mixture HMM's developed recently [12] in that the unimodal multivariate Gaussian densities in our models are distinct not only for each mixture component but also for each state transition (see (1)).

TABLE IV
EFFECTS OF DIFFERENT NUMBERS OF MIXTURE COMPONENTS ON RECOGNITION PERFORMANCE. UNIFORM LANGUAGE MODEL

Number of Mixture Components	Percent Correct (Speaker AM)	Percent Correct (Speaker CA)
1	68.7%	70.6%
5	73.2%	79.4%
25	76.6%	86.1%
38	75.0%	86.0%

When evaluated in an 86 000-word recognizer with a trigram language model, we found that 25-component mixture phonemic HMM's significantly outperform unimodal phonemic HMM's, leading to a reduction of more than 40% in the error rate when the training set is larger than about 2500 words. We also found that under the same conditions, 25-component mixture models gave significantly better results than a set of triphone HMM's consisting of 25 unimodal models per phoneme. The is rather surprising since it seems to suggest that our recognizer performs better without information concerning phonetic context. Perhaps a more reasonable explanation is that context-dependent allophone models would perform better if they too were modeled using mixture HMM's. Unfortunately, it is difficult to see how this can be done in practice as there are more than 17 000 triphones in our dictionary and schemes for tying triphone models based on prior phonetic knowledge are unsatisfactory. Our results suggest that in training very large vocabulary speech recognizers with moderate amounts of data (2000 to 3000 words), the free parameters are better used to construct large mixture models for phonemes rather than attempting to explicitly model context dependence.

REFERENCES

[1] R. K. Moore, M. J. Russell, and M. J. Tomlinson, "The discriminative network: A mechanism for focusing recognition in whole-word pattern matching," in *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 1041–1044, 1983.

[2] V. Gupta, M. Lenning, and P. Mermelstein, "Fast search strategy in a large vocabulary word recognizer," *J. Acoust. Soc. Amer.*, vol. 84, pp. 2007–2017, 1988.

[3] L. Deng, M. Lennig, P. Kenny, and P. Mermelstein, "Modeling acoustic transitions in speech by state-interpolation hidden Markov models," *IEEE Trans. Signal Processing*, to be published.

[4] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speaker independent isolated word recognition," in *Proc. ICASSP*, 1986, pp. 41–44.

[5] A. B. Poritz and A. G. Richter, "On hidden Markov models in isolated word recognition," in *Proc. ICASSP*, 1986, pp. 705–708.

[6] A. Nadas and D. Nahamoo, "Automatic speech recognition via pseudoindependent marginal mixtures," in *Proc. ICASSP*, 1987, pp. 1285–1287.

[7] H. Ney and A. Noll, "Phoneme modeling using continuous mixture densities," in *Proc. ICASSP*, 1988, pp. 437–440.

[8] B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 307–309, 1986.

[9] L. Deng, M. Lennig, and P. Mermelstein, "Use of vowel duration information in a large vocabulary word recognizer," *J. Acoust. Soc. Amer.*, vol. 86, pp. 540–548, Aug. 1989.

[10] L. Deng, M. Lennig, and P. Mermelstein, "Modeling microsegments

of stop consonants for speech recognition,'' *J. Acoust. Soc. Amer.*, vol. 87, pp. 2738–2747, June 1990.

[11] L. Deng, M. Lennig, F. Seitz, and P. Mermelstein, ''Large vocabulary word recognition using context-dependent allophonic hidden Markov models,'' *Comput. Speech Language*, vol. 4, no. 2, pp. 193–202, 1990.

[12] J. R. Bellegarda and D. Nahamoo, ''Tied mixture continuous parameter modeling for speech recognition,'' *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 12, pp. 2033–2045, Dec. 1990.