Acoustic Recognition Component of an 86,000-word Speech Recognizer*

L. Deng[1], V. Gupta[2], M. Lennig[2], P. Kenny, and P. Mermelstein[2]
INRS-Télécommunications, Montreal, Quebec, Canada

Abstract

This paper describes recent results obtained with an HMM-based acoustic recognizer using a virtually unlimited vocabulary (86,000 words) to perform speaker-dependent isolated-word recognition. The task domain of this recognizer is quite general, consisting of paragraphs read from various newspapers, books and magazines. We present the results of a comparative acoustic recognition study using various types of HMM's and various amounts of training data (from 700 to about 4000 words). The models explored include context-dependent allophonic HMM's (including generalized diphone and triphone models with unimodal Gaussian output densities) and context-independent phonemic HMM's (using either unimodal or mixture densities). Experimental results indicate that phonemic HMM's with many components in the mixture output densities provide the highest acoustic recognition accuracy. The acoustic recognition accuracy for a total of about 7000 test words spoken by 4 male and 5 female speakers is 82%. Recognition accuracy after application of the language model increases to 92%.

I. Introduction

We have previously evaluated the performance of a large vocabulary speaker dependent isolated word recognizer where each phoneme is represented by one HMM with unimodal Gaussian output densities (Gupta et al. 1987, 1988; Deng et al. 1988a,b; Deng et al. 1989a,b). The acoustic recognition accuracy achievable by that system is severely limited no matter how much training data is used. The major problem considered responsible for the poor recognition accuracy is the high variability in the acoustic realization of a phoneme (especially near the boundaries) arising from the variety of phonemic contexts it occurs in. Much higher variances are observable in the Gaussian distributions associated with the HMM states near the phoneme boundaries than with those near the center (Deng et al., 1989a).

Two strategies are available to overcome this problem. The first is to condition a unimodal HMM on the specific phonemic context, resulting in a set of context-dependent unimodal HMMs. In the second strategy, we improve the modeling of contextual variation by extending the unimodal output densities to Gaussian mixtures. We expect that the self-organization capability of this more general model will automatically capture the acoustic variability via the multiple modes in the Gaussian mixture distribution. While phonemes represented by the mixture HMM's described here are context-independent, the multimodal output densities can be expected to provide a more complete representation of the observed acoustic parameter variation in the initial and final HMM states than the unimodal Gaussian output densities. In this paper, we report experimental results obtained using the above strategies for overcoming the acoustic variation problem.

This paper is organized as follows. In Section II, we describe the construction of the context-dependent HMM, emphasizing techniques used for reliable estimation of the model parameters. In Section III, we give a general description of the mixture HMM used in this work and explain how it can be easily implemented as a special type of unimodal Gaussian HMM which we call a multiple-branch HMM. The experimental evaluation of the phonemic mixture HMM's and the context-dependent allophonic HMM's in an 86,000-word recognizer is presented in Section IV. Finally, we summarize our results in Section V.

II. context-dependent allophonic HMMs

Before describing the context-dependent allophonic HMM's, we first review the context-independent phonemic HMM's implemented in the baseline recognition system developed previously (Gupta et al., 1988; Deng et al., 1988b). Briefly, the HMM representing a phoneme is based on an underlying left-to-right Markov chain starting in state 1 and ending in state N. N ranges from 4 to 10 for the different phonemes. Transitions between phonemes are represented by a null transition from the last state of one phonemic HMM to the first state of the next phonemic HMM. With each transition in the Markov chain, we associate a unimodal multivariate Gaussian distribution of speech feature vectors. A context-independent HMM for a phoneme is trained using all tokens of this phoneme and invoked to evaluate any instance of this phoneme during recognition. Although such context-independent HMMs can be trained reliably, the discrimination they provide is often inadequate because of the well known co-articulation effects in speech.

* This work was supported by the Natural Sciences and Engineering Research Council of Canada

[1] Currently with Electrical Engineering Department, University of Waterloo, Ontario, Canada

[2] Also with BNR, Montreal, Canada

CH2847-2/90/0000-0741 $1.00 © 1990 IEEE

Context-dependent allophonic HMMs, which model a phoneme in a particular context, are designed to circumvent this inadequacy. In this study, a context refers to the immediate left and/or right adjacent phonemes. A one-sided allophonic HMM for a phoneme is the allophonic HMM trained with the tokens of this phoneme conditioned on the left *or* the right adjacent phoneme, and invoked for recognition only when this phoneme occurs in the given context. By this definition, two allophones of the same phoneme having different left or right context are distinct. We use the notation l/r-allophone to denote the one-sided allophone conditioned on either the *left* or the *right* context.

Likewise, a two-sided allophonic HMM is the allophonic HMM having the context of the phoneme defined as both the left *and* the right adjacent phonemes. The two-sided allophone is denoted as an lr-allophone.

The allophonic HMM's defined above are powerful representational tools since they capture the most direct coarticulation effects in speech. However, many of these allophonic models, especially the two-sided ones, would suffer from lack of sufficient training data. Introducing additional allophones reduces the number of training tokens available for each allophone.

One approach to overcoming this difficulty is based on the observation that certain groups of phonemes have similar co-articulatory effects. For instance, consonants sharing the same place of articulation tend to affect the adjacent vowel in a similar way. Also, vowels pronounced with a similar tongue position tend to have similar effects on neighboring consonants. Systematic application of such speech knowledge allows construction of a set of merged contexts which are phonetically coherent. The selected contextual groups are shown below.

The contexts affecting vowels are: (1) word boundary, breath, and /h/; (2) labial consonants; (3) dental, alveolar and palatal consonants; (4) velar consonants; (5) another vowel. The contexts for a consonant are: (1) word boundary, breath; (2) palatal vowels, /j/; (3) rounded vowels, /w/; (4) plain vowels; (5) another consonant, except /j/ or /w/.

To construct one-sided HMM's using these merged contexts, we use the preceding phoneme class for a vowel and the following phoneme class for a consonant. In this way, each phoneme is represented by five context-dependent one-sided allophonic HMM's. Note that due to this directional definition of context dependence, a diphthong can be easily treated as one of the vowel categories when encountered as a context. In constructing two-sided allophonic HMM's for each phoneme (vowel or consonant), a combination of the above five contexts in both left *and* right gives rise to 25 two-sided allophonic contexts. After the merger of contexts, we denote one-sided and two-sided allophones as L/R-allophones and LR-allophones, respectively.

Despite the reduction in contextual categories described above, there are still numerous allophonic HMM's to be trained for each phoneme. This is especially true for the LR-allophonic HMM's. Hence it is desirable to employ smoothing techniques to avoid possibly unreliable estimates of the HMM parameters. Two such techniques have been used. First, the output covariance matrices of all allophonic HMM's for a phoneme are tied and trained as a single covariance matrix of the phoneme. Because Gaussian HMM's are in general more sensitive to estimation errors of the output means than to those of the covariance matrices, making the covariance matrix independent of context is not expected to degrade the ability of the model to represent context-dependent effects. Also, tying naturally leads to estimation of covariance matrices in the allophonic HMM which are as reliable as those in phonemic HMM's. Second, to insure reliable estimation of the Gaussian mean parameters in the allophonic HMM's, they are interpolated with those in the corresponding phonemic HMM's.

III. Mixture HMM's

The parameters that characterize the mixture HMM are: (1) $A = [a_{ij}], i, j = 1, 2, \ldots, N$, is 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 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

$$b(\mathbf{O}) = \sum_{m=1}^{M} c_m g_m(\mathbf{O}) \qquad (1)$$

where $\mathbf{O}$ is the observation vector, and for each mixture component $m = 1, \ldots, M$, c_m is the mixture weight ($\sum_{m=1}^{M} c_m = 1$) and $g_m(\mathbf{O})$ is a unimodal multivariate Gaussian density.

The training algorithm for mixture HMM's is given by Juang et al. (1986). However, we have 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 (Liporace, 1982), 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. Associated with each transition in the mixture HMM there is a transition probability p, and, for each $m = 1, \ldots, M$, a mixture weight c_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_m and whose output distributions are the corresponding mixture components. We will refer to this unimodal model as a *multiple-branch* HMM. By adopting the equivalence of the mixture HMM and the multiple-branch HMM, our previous software for the Baum-Welch training of the unimodal HMMs is easily extended to train the mixture HMMs.

IV. Speech recognition experiments

A. Introduction

This section describes a series of experiments to evaluate the effectiveness of the context-dependent allophonic HMM's and the phonemic mixture HMM's in a speaker-dependent isolated-word recognition task using 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 25.6 ms Hamming window is applied at intervals of 10 ms, and each frame is represented by a 15-dimensional vector consisting of mel-frequency cepstral coefficients (Davis and Mermelstein, 1980) and their differences over time. The data consist of English passages spoken as sequences of isolated words. These passages were selected arbitrarily from magazines, books, office correspondence and newspapers. The number of words spoken by each speaker varies between 2,000 and 5,000.

The structure of our large-vocabulary isolated-word recognizer has been described previously (Gupta et al., 1988; Deng et al., 1988b). 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 the uniform language model[3] or the trigram language model trained on 60-million words of text.

B. Results with 25-component mixture HMMs

Recognition results from nine speakers are presented in Table 1. For each speaker, the number of words used in training and test sets, and the recognition accuracy achieved at various stages of the mixture-HMM recognizer are shown in separate columns. The test data comprises 6,816 words.

Shown in the *search* column is the performance of the fast search strategy based on the syllabic graph search algorithm described by Gupta et al. (1988). This column shows the percentage of test words selected as one of 300 candidate words from among 86,000 words in the vocabulary. When averaged over the nine speakers, the fast search algorithm misses 3.0% of the words. This result, using mixture HMM's in the fast search, is significantly better than that using unimodal Gaussian HMM's.

Shown in the *uniform* and *trigram* columns are the recognition accuracies using the uniform and the trigram language models. Accuracy is measured by the percentage of test words recognized as the top word choice. Homophone confusions *are* counted as errors when the trigram language model is used, but *are not* when the uniform language model is used. The weighted averages of the recognition rates for the uniform and the trigram language models are 82.2% and 91.8% respectively. This result represents our highest recognition accuracy so far.

[3] I.e., all words are considered a priori equally probable. The recognition accuracy is purely due to the acoustic information in the speech.

C. Context-dependent allophonic HMM's versus phonemic unimodal and mixture HMM's

Table II shows the recognition accuracy obtained by the context-dependent allophonic HMM's (headings *L/R-alloph.* and *LR-alloph.*), in comparison with the benchmark (context-independent unimodal phonemic HMM's, heading *benchmark*) and with the phonemic mixture HMM's (heading *mixtures*). The benchmark system with phonemic HMM's has been described previously (Gupta et al., 1988; Deng et al., 1988b). The L/R-allophonic, LR-allophonic and mixture HMM's are described in the preceding sections of this paper. The first conclusion we can draw from Table II is that allophonic HMM's consistently outperform unimodal phonemic HMM's. The difference in recognition accuracy is particularly noticeable with a large amount of training data (e.g. over 2,500 words). In this case, averaged over speakers CA and AM, L/R-allophonic HMM's reduce recognition errors by 18% when we use the uniform language model and by 26% when we use the trigram language model. LR-allophonic HMM's reduce the error rate further, by 35% and 33%, respectively, for the two language models.

For speakers CA and AM, varying amounts of training data have been used in an attempt to investigate the effects of the extent of context coverage between the training and test data on the recognition accuracy. The context coverage is mainly determined by the training size (Deng et al., 1989c). The effects, demonstrated by the results in Table II, can be summarized as follows. An increase in the training size and consequently in the degree of context coverage improves the recognition accuracy for both phonemic and allophonic HMM's. The improvement saturates beyond a certain size of training data. The saturation point is the lowest for unimodal phonemic HMM's, intermediate for L/R-allophonic HMM's, and the highest for LR-allophonic HMM's.

It can be seen from Table II that in some cases L/R-allophonic HMM's produce higher recognition accuracy than LR-allophonic HMM's. This is unexpected since LR-allophonic HMM's in general capture more detailed co-articulatory information than do L/R-allophonic HMM's. We suspect that the poorer recognition accuracy for LR-allophonic HMM's results from unreliable estimation of mean vectors in certain models, since there are four times more of such mean vectors than in an L/R-allophonic HMM. This conjecture is consistent with the fact that LR-allophonic HMM's perform worse than L/R-allophonic HMM's when the trigram language model is used, although the opposite is true with the uniform language model (see the results for speaker MA, and for speaker AM with 2742 training words in Table II).

The results in Table II with mixture HMM's were obtained as follows. We used the mean vectors from the LR allophonic HMM's as an initialization for training mixture HMM's (one per phoneme). The two sets of HMMs thus have the same number of parameters. It came as something of a surprise to discover that the mixture HMM's outperformed the LR allophonic HMM's in almost every instance

both with the uniform and the trigram language models (compare columns 7–8 to columns 9–10).

V. Summary and conclusion

Two different approaches are described in this paper to deal with the problem of acoustic variation of phonemes encountered in varying phonetic contexts in a large vocabulary speech recognition system. First, we constructed context-dependent allophonic HMM's. We then implemented mixture HMM's in the unimodal transition–based framework by 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 main conclusions of the study are as follows. First, successful modeling of context-dependent allophones requires reliable estimation of allophone model parameters. Reliable estimates of the allophone model parameters are obtained by careful selection of the allophones, and by tying the covariance matrix for the phoneme across all the HMM state transitions in all the allophone models of the phoneme. The allophone models for the phoneme are selected by grouping the neighboring phonemes into classes based on their expected coarticulatory effects on the phoneme being modeled. Second, both the phonemic mixture HMM's and the allophonic unimodal HMM's consistently outperform the phonemic unimodal HMM's for almost any amount, but more significantly for a large amount of training data. Third, the allophonic unimodal HMM's are inferior to the phonemic mixture HMM's, despite the use of contextual information and of the same number of parameters in the two types of models. This 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 modelled using mixture HMM's. Unfortunately, it is difficult to see how this can be done in practice as more than 17,000 l/r allophones are counted in our dictionary and schemes for reducing this number by constructing L/R allophones based on prior phonetic knowledge proposed in this paper are still unsatisfactory. Our results suggest that in training very large vocabulary speech recognizers with moderate amounts of data, the free parameters are better used to construct large mixture models for phonemes instead of attempting to model context–dependence explicitly.

Speaker (sex)	Training size	Test size	recognition acuracy		
			search	uniform	trigram
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%
Avg.	2081	6816	97.0%	82.2%	91.8%

Table I. Recognition results using 25-component mixture HMM's for nine speakers.

Speaker *(test size)*	Train size	Benchmark		L/R-alloph.		LR-alloph.		Mixtures	
		unif.	trig.	unif.	trig.	unif.	trig.	unif.	trig.
CA (female) (1090 words)	717	67.9	80.6	70.0	83.5	69.7	82.5	69.7	80.5
	1532	70.2	85.0	75.0	88.5	78.1	88.0	81.0	90.0
	2347	70.6	86.8	75.5	89.1	80.8	90.4	86.1	94.2
	3098	70.8	86.8	76.0	89.0	82.7	91.3	86.0	94.0
	3880	70.7	86.7	76.1	89.1	83.0	91.9	85.9	94.0
AM(male) (698 wds)	1100	54.4	78.0	54.0	78.0	53.4	77.0	55.4	79.0
	2039	68.2	81.0	73.0	84.0	72.0	83.0	73.9	87.0
	2742	68.7	81.9	74.2	88.1	76.1	86.4	76.7	89.7
MA(fem.) (586 wds)	1600	79.0	90.4	83.1	91.6	83.8	89.6	86.2	92.5

Table II. Comparison of recognition accuracy (in %) for the context-dependent allophonic HMM's (L/R diphone and LR triphone models) and the context-independent phonemic HMM's (unimodal and mixture models). Results for the uniform (unif.) and trigram (trig.) language models are given separately.

References

S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–365, 1980.

L. Deng, M. Lennig, V. Gupta, and P. Mermelstein "Modeling acoustic-phonetic detail in an HMM-based large vocabulary speech recognizer," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 509–512, 1988a.

L. Deng, P. Kenny, M. Lennig, V. Gupta and P. Mermelstein, "Large vocabulary word recognition based on phonetic representation by hidden Markov models", *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, pp. 131-134, 1988b.

L. Deng, P. Kenny, M. Lennig, and P. Mermelstein. "Modeling acoustic transitions in speech by state-interpolation hidden Markov models," to appear in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1989a.

L. Deng, M. Lennig, and P. Mermelstein. "Use of vowel duration information in a large vocabulary word recognizer," *Journal of the Acoustical Society of America*, Vol. 86, pp. 540–548, 1989b.

L. Deng, M. Lennig, F. Seitz and P. Mermelstein. "Large vocabulary word recognition using context-dependent allophonic hidden Markov models," submitted to *Computer Speech and Language*, 1989c.

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

V. Gupta, M. Lennig and P. Mermelstein, "Fast search strategy in a large vocabulary word recognizer," *Journal of the Acoustical Society of America*, vol. 84, pp. 2007–2017, 1988.

B.H. Juang, S.E. Levinson and M.M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Transactions on Information Theory*, vol. IT-32, pp. 307-309, 1986.

L.A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Transactions on Information Theory*, vol. IT-28, pp. 729-734, 1982.