Computer Speech and Language (1990) **4**, 345–357

Large vocabulary word recognition using context-dependent allophonic hidden Markov models*

L. Deng, M. Lennig, F. Seitz and P. Mermelstein

INRS-Télécommunications, Montreal, Quebec, H3E 1H6, Canada

Abstract

In this paper, we report our development of context-dependent allophonic hidden Markov models (HMMs) implemented in a 75 000-word speaker-dependent Gaussian-HMM recognizer. The context explored is the immediate left and/or right adjacent phoneme. To achieve reliable estimation of the model parameters, phonemes are grouped into classes based on their expected co-articulatory effects on neighboring phonemes. Only five separate preceding and following contexts are identified explicitly for each phoneme. By grouping the contexts we ensure that they occur frequently enough in the training data to allow reliable estimation of the parameters of the HMM representing the context-dependent units. Further improvement in the estimation reliability is obtained by tying the covariance matrices in the HMM output distributions across all contexts. Speech recognition experiments show that when a large amount of data (e.g. over 2500 words) is used to train context-dependent HMMs, the word recognition error rate is reduced by 33%, compared with the context-independent HMMs. For smaller amounts of training data the error reduction becomes less significant.

1. Introduction

Currently, most successful speech recognizers have been built using hidden Markov models (HMMs) to represent words (Lippmann, Martin & Paul, 1987; Rabiner, Wilpon & Soong, 1988). While remarkable performance has been demonstrated in small vocabulary tasks, this word-model based approach cannot be straightforwardly extrapolated to very large vocabulary applications. The main reason for this is that in practice the HMMs representing words cannot be adequately trained with a limited amount of training data.

An alternative to the word-model based approach is to use speech units smaller than words. There are two major classes of this subword modeling approach, one based on the traditional concept of phoneme and the other one based on acoustic segments. Acoustic segment-based HMM systems (Bahl *et al.*, 1988; Lee, Soong & Juang, 1988)

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

0885–2308/90/040345 + 13 $03.00/0

avoid any preconception of traditional phonetic units, and construct the basic recognition units entirely on the basis of the acoustic similarity. The units chosen in this way are acoustically consistent and typically very small and numerous. Thus the influence of one unit on another is reduced to a minimum. However, one serious drawback of this approach is the requirement to construct the acoustic lexicon by defining a new baseform for each word in the vocabulary. This drawback argues against the use of an acoustic segment-based system for very large vocabulary recognition.

On the other hand, the phonemic HMM system relies upon the linguistic notion of phonemes corresponding to the segments serving as the minimal differentiators between words. Although most problems in word-based systems and acoustic segment-based subword systems are eliminated, performance of phoneme-based systems has been relatively poor if these segments are considered invariant (Bahl *et al.*, 1980; Merialdo, 1987; Deng *et al.*, 1988*a*; Gupta, Lennig & Mermelstein, 1988; Lee & Hon, 1988; Murveit & Weintraub, 1988). The reason why phonemic models are inadequate is that they assume a phoneme in any context is equivalent to the same phoneme in other contexts. However, due to coarticulation (i.e. articulators cannot move instantaneously from one configuration to another), the acoustic realization of a phoneme is strongly affected by others, especially by its immediate neighbors.

To capture coarticulatory effects, context-dependent allophonic HMMs have been proposed and implemented in several HMM-based recognition systems (Schwartz *et al.*, 1984; Chow *et al.*, 1987; Lee, 1988; Paul & Martin, 1988). In these systems, the most successful type of context-dependent allophonic HMMs is the phonemic model which takes into account the immediately left and right phonemic contexts. However, since this type of method for determining HMM contexts relies upon enumeration of contexts, the reliability of parameter estimation of the resulting context-dependent HMMs, and their effectiveness in speech recognition, become inherently limited by the overlap of common contexts between the training and test data. This overlap is determined principally by the amount of training data (assuming that training texts are selected with no consideration of phonetic balance) and by the size of the vocabulary.

Context-dependent allophonic HMM recognizers developed previously typically have vocabularies consisting of no more than 1000 words. As such, even a small amount of training data can easily provide sufficient context coverage in the test set (Chow *et al.*, 1987; Lee, 1988; Paul & Martin, 1988). This clearly accouts for the success of the context-dependent allophonic HMMs implemented in these recognizers. However, for HMM recognizers with much larger vocabularies, the training set typically manifests only a fraction of the phonological contexts that occur in the test data. (The statistics we collected which will be shown in Section 4 indicates that about one-half of triphones in test data are not observable in a set of typical training data containing as many as 1500 words.) Due mainly to this reason, the 10 000-word HMM recognizer reported by Derouault (1987) showed somewhat discouraging results on the use of the enumeration-based context-dependent allophonic HMMs (only 9·3% word error reduction).

In this paper, we report encouraging results we obtained from context-dependent allophonic HMMs implemented in a 75 000-word Gaussian HMM recognizer. The significant improvement we have achieved in recognition accuracy over the context-independent HMM recognizer is attributed to two techniques we have developed and employed. First, in defining allophonic HMMs, we merge similar contexts together according to acoustic-phonetic knowledge. This context merging results in a great reduction in the number of models to be trained, and consequently a large increase in the

training data for each model. The second technique is to impose the constraint that the covariance matrices of allophonic HMMs are identical (tying) for all contexts of the phoneme. The tying makes the estimation of the covariance matrix more reliable, and significantly reduces the amount of training data required.

This paper is organized as follows. In Section 2, we present the context-dependent allophonic units developed and employed in this work. In Section 3, we describe the Baum–Welch training procedure employed to construct the HMMs representing these units. An analysis on the overlap of common contexts between training and test data is shown in Section 4. The result of this analysis suggests a strong need for merging similar contexts. In Section 5, the effectiveness of the context-dependent allophonic HMMs is demonstrated in comparison with context-independent phonemic HMMs in a 75 000-word speaker-dependent isolated-word recognizer. Finally, we discuss and summarize our results in Section 6.

2. Hidden Markov models representing phonemes and allophones

Before describing the context-dependent allophonic HMMs, we first review the context-independent phonemic HMMs implemented in the baseline recognition system developed previously (Deng *et al.*, 1988*a*; Gupta, Lennig & Mermelstein, 1988). 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 another 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 with all tokens of this phoneme and invoked to evaluate any instance of this phoneme during recognition. Although such a context-independent HMM can be trained reliably due to its frequent occurrence in the training data, it is often inadequate because of the well-known coarticulation effects in speech.

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 same context. By this definition, two allophones of the same phoneme having different left or right context are considered distinct. For notational convenience, we use 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 lr-allophone.

The allophonic HMMs defined above are powerful representational tools since they capture the most direct coarticulation effects in speech. However, if the left or/and the right context for these models is taken to be a specific phoneme, many of these models, especially the two-sided HMMs, would be poorly trained. By allowing many more models, the number of occurrences of the allophone in the training data used to train each model is reduced.

One approach to overcoming this difficulty is based on the observation that some phonemes have very similar effects on the adjacent phonemes. For instance, consonants

sharing a place of articulation feature tend to affect the successive vowel in a similar way; and vowels pronounced with a similar tongue position tend to have similar effects on left-neighboring consonants. A systematic application of this type of knowledge allows construction of a set of merged contexts which are phonetically meaningful. The resulting contexts of a phoneme after the context merging are shown below.

The merged contexts for a vowel are: (1) word boundary, breath, or /h/; (2) labial consonants; (3) dental, alveolar, and palatal consonants; (4) velar consonants; (5) another vowel.

The merged contexts of 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 HMMs 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 HMMs. 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 HMMs 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 use L/R-allophone and LR-allophone to denote one-sided and two-sided allophones, respectively.

Despite the reduction in contextual categories described above, there are still numerous allophonic HMMs to be trained for each phoneme. This is especially true for the LR-allophonic HMMs, 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 LR-allophonic or L/R-allophonic HMMs for a phoneme are tied and trained as a single covariance matrix of the phoneme. Because Gaussian HMMs are in general more sensitive to estimation errors of the output means than to those of the covariance matrices (Rabiner *et al.*, 1985), making the covariance matrix independent of context is not expected to degrade the ability of the model to represent context-dependent effect. Nevertheless, tying naturally leads to estimation of covariance matrices in the allophonic HMM which are as reliable as those in phonemic HMM. Second, to insure reliable estimation of the Gaussian mean parameters in the allophonic HMMs, they are interpolated with those in the phonemic HMM which have been reliably trained. In the next section, we describe implementation details of these smoothing techniques, together with the general procedure for training allophonic HMMs.

3. Training allophonic hidden Markov models

A set of C context-dependent Gaussian HMMs are used to represent C context-dependent allophones of a phoneme. In the present study, C is 5 for L/R-allophonic context, and 25 for LR-allophonic context. These HMMs are characterized by the following parameters:

(1) $[a_{ij}^{(c)}]$, $i,j = 1,2,\ldots,N$ and $c = 1,\ldots,C$, a set of state transition matrices of the Markov chains each with N states. Here $a_{ij}^{(c)}$ is the transition probability from state i to state j for the specific context c. For the left-to-right HMM, we assume $a_{ij}^{(c)} = 0$, for $j < i$ and $j > i+2$.

(2) An output probability density for the observation vector defined on each state transition in the Markov chain for each specific context. The density is assumed to be a multivariate Gaussian of the form:

$$b_{ij}^{(c)}(\mathbf{O}) = N[\mathbf{O}, \Theta_{ij}^{(c)}, \Sigma]$$

where $\mathbf{O}$ is the observation vector, $\Theta_{ij}^{(c)}$ is the mean vector distinct for each transition from state i to j, and specific for each context c. The covariance matrix Σ, is assumed to depend only on the phoneme, common to all allophones of that phoneme and to all state transitions.

The above parameters are estimated from training data by the well-known Baum–Welch re-estimation algorithm, an iterative procedure for maximum likelihood estimation of HMM parameters (Baum, 1972). Since the re-estimation formulas and their derivation for context-dependent Gaussian HMMs with tied parameters are still quite similar to those for standard Gaussian HMMs (Liporace, 1982), we only give the final formulas and note the differences.

A transition probability from state i to j in the HMM representing a phoneme in context c is re-estimated by

$$a_{ij}^{(c)} = \frac{\sum_{k=1}^{K_c} \sum_{t=1}^{T_c^{(k)}} \gamma_t^k(i,j,c)}{\sum_{k=1}^{K_c} \sum_{t=1}^{T_c^{(k)}} \sum_{j=1}^{N} \gamma_t^k(i,j,c)}, \tag{1}$$

where $T_c^{(k)}$ is the total number of observation vectors in token k of allophone c, K_c is the total number of tokens of this allophone, and

$$\gamma_t^k(i,j,c) = \frac{P(\mathbf{O}_1^{(k,c)}, \mathbf{O}_2^{(k,c)}, \ldots, \mathbf{O}_{Tc(k)}^{(k,c)}, s_{t-1}=i, s_t=j)}{P(\mathbf{O}_1^{(k,c)}, \mathbf{O}_2^{(k,c)}, \ldots, \mathbf{O}_{Tc(k)}^{(k,c)})} \tag{2}$$

is the conditional probability that for token k of allophone c, a state transition from i to j takes place at time $t-1$, given that the observation sequence is generated by the model.

The re-estimation formula for the mean vector of the Gaussian density in the HMM representing allophone c is

$$\Theta_{ij}^{(c)} = \frac{\sum_{k=1}^{K_c} \sum_{t=1}^{T_c^{(k)}} \gamma_t^k(i,j,c) \mathbf{O}_t^{(k,c)}}{\sum_{k=1}^{K_c} \sum_{t=1}^{T_c^{(k)}} \gamma_t^k(i,j,c)}, \tag{3}$$

and the re-estimate of the covariance matrix is

$$\sum = \frac{\sum_{k=1}^{K_c} \sum_{t=1}^{T_c^{(k)}} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{c=1}^{C} \gamma_t^k(i,j,c)(\mathbf{O}_t^{(k,c)} - \Theta_{ij}^{(c)})(\mathbf{O}_t^{(k,c)} - \Theta_{ij}^{(c)})^*}{\sum_{k=1}^{K_c} \sum_{t=1}^{T_c^{(k)}} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{c=1}^{C} \gamma_t^k(i,j,c)}. \tag{4}$$

Note that the tying of the covariance matrices across all contexts and all state transitions is carried out by summing up the accumulated quantities in both numerator and denominator over i, j and c in Equation (4).

The conditional probability $\gamma_t^k(i,j,c)$ [defined by Equation (2)] which is involved in Equations (1), (3) and (4) can be efficiently calculated by using standard forward and backward probabilities (Baum, 1972; Bahl, Jelinek & Mercer, 1983; Levinson, Rabiner & Sondhi, 1983).

The Baum–Welch re-estimation formulas given above assume that the phonetic segment boundaries are known for each training token. In practice, however, the phonetic segment boundaries are often quite inaccurate. To overcome this difficulty, an automatic training procedure is required which allows adjustments of the phoneme segment boundaries. We implemented such a procedure which considers all possible phonetic segments in a word. The procedure is efficiently implemented by concatenating a series of allophonic HMMs, which correspond to the phonemic transcription of the word and respect their immediate left and/or right phonetic contexts, as a training token. The HMM network constructed in this way to represent words from allophones can then be trained by directly applying the re-estimation formulas in Equations (1), (3) and (4). When common allophones are present in a word, the HMM parameters of the common allophones are tied in applying the re-estimation formulas.

As explained before, we obtain reliable estimates for the covariance matrices by tying them across all contexts for a given phoneme. To obtain reliable estimates for the mean vectors, we used the following formula to linearly interpolate the mean vectors re-estimated at the m-th iteration with those at the previous iteration:

$$\Theta^{(m)} \leftarrow \lambda\Theta^{(m)} + (1-\lambda)\Theta^{(m-1)}, \tag{5}$$

where the interpolation weight λ is a function of the expected number of frames (determined by the Baum–Welch algorithm) used to estimate the $\Theta^{(m)}$. Only a fixed number of iterations were carried out, which was typically long before the convergence of the Baum–Welch algorithm was reached. Baum and Sell (1968) showed that such an interpolation operation retains the essential property of the Baum–Welch algorithm that each iteration guarantees an increase in the likelihood on the training data. In all the experiments reported in this paper, context-independent HMMs are used as initial models in training context-dependent HMMs. Therefore, if tokens for a particular context are not present or occur very rarely, then Equation (5) essentially becomes a smoothing of the re-estimate of a context-dependent mean vector by that of a context-independent mean vector.

4. Analysis of context coverage

As pointed out earlier, for the enumeration-based approach to context-dependent phonetic modeling which we explore here, the overlap of common contexts between the training and test sets is a major factor determining the success of the approach. Before we present experimental evaluation of the context-dependent allophonic HMMs, we show in this section the results of an analysis on the context coverage. The purpose of this analysis is to demonstrate the necessity of merging the allophonic contexts as described in Section 2, and to allow for quantitative interpretations of the dependence of recognition accuracy on the training size.

The texts employed to carry out the context coverage analysis are paragraphs randomly sampled from four novels. The training and test data are disjoint, and consist of 3880 and 1090 words, respectively.

Table I shows quantitative measures of the allophone coverage between the 1090-word test set and the training sets comprising 717, 1532, 2343, 3098 and 3880 words, respectively. The l/r-allophone and lr-allophone discussed here are those defined in Section 2 before context merging is considered. The context coverage is measured by the percentage of distinct l/r-allophones or lr-allophones in the test set which are present with various occurrence frequencies in the training set. The total number of distinct l/r-allophones or lr-allophones in the 1090-word test set is 612 and 1500, respectively. The percentage values outside the parentheses are those measures which do not take into account the actual occurrence frequencies of the allophones in the test data (they simply provide the weight one to an allophone if the allophone is in the test data and provide the weight zero if the allophone is not in the test data), while the percentage values inside the parentheses are those which weight each allophone by such frequencies. Both sets of values are meaningful since in actual speech recognition experiments not only the context-dependent model representing the test word, but also the models representing practically every word in the vocabulary, are to be used to match the unknown input acoustic observations. Therefore, although the degree of context overlap appears significantly higher by the measure where the allophone occurrence frequency weighting is used, the effect of such a higher degree of context overlap would not totally manifest itself in speech recognition experiments.

As can be seen from the table, a large percentage of allophones in the test set occur with very low frequencies even in large training sets. This is especially true for lr-allophones. Apparently, if distinct HMMs were used to represent all of these allophones, their parameters could not be reliably estimated.

This serious problem motivates the merging of contexts. Table II demonstrates a significantly improved context coverage between training and test sets for L/R-allophones or LR-allophones after the context merging described in Section 2. For

TABLE I. Percentage of distinct one-sided and two-sided allophones (both before context merge) in a 1090-word test set occurring with various frequencies in disjoint training sets with varying size. The percentage values in the parentheses are obtained with the weighting of each allophone by the number of its occurrences in the test set

Type of allophones	Occurrences in training data	Training size (No. of words)				
		717	1532	2343	3098	3880
l/r-allophones	0	27% (7%)	14% (3%)	9% (1·3%)	7% (1·1%)	5% (0·8%)
	1	17% (1·0%)	12% (0·8%)	11% (0·6%)	8% (0·3%)	6% (0·3%)
	2 to 5	34% (15%)	32% (8%)	25% (5%)	22% (4%)	19% (3%)
	6 to 10	10% (15%)	19% (11%)	17% (8%)	20% (5%)	21% (4%)
	more than 10	12% (61%)	23% (78%)	37% (85%)	43% (90%)	49% (92%)
lr-allophones	0	56% (23%)	41% (15%)	33% (11%)	28% (9%)	23% (7%)
	1	20% (12%)	19% (8%)	19% (7%)	16% (6%)	17% (6%)
	2 to 5	18% (19%)	27% (16%)	29% (15%)	30% (15%)	30% (15%)
	6 to 10	3% (10%)	7% (8%)	10% (9%)	13% (9%)	14% (9%)
	more than 10	3% (36%)	6% (53%)	9% (58%)	12% (61%)	16% (63%)

TABLE II. Percentage of distinct one-sided and two-sided allophones (both after context merge) in a 1090-word test set occurring with various frequencies in disjoint training sets with varying size. The percentage values in the parentheses are obtained with the weighting of each context-merged allophone by the number of its occurrences in the test set

Type of allophones	Occurrences in training data	Training size (No. of words)				
		717	1532	2343	3098	3880
L/R-allophones	0	7% (0·4%)	3% (0·3%)	2% (0·2%)	2% (0·2%)	1% (0·1%)
	1	7% (1·3%)	1% (0·9%)	2% (0·4%)	2% (0·2%)	1% (0·1%)
	2 to 5	27% (4%)	17% (3%)	9% (1·3%)	5% (0·9%)	4% (0·9%)
	6 to 10	14% (9%)	12% (12%)	12% (4%)	12% (4%)	9% (3%)
	more than 10	44% (85%)	66% (85%)	75% (95%)	80% (95%)	85% (96%)
LR-allophones	0	20% (4%)	11% (1·0%)	8% (0·5%)	6% (0·4%)	4% (0·4%)
	1	18% (5%)	9% (2%)	7% (0·5%)	6% (0·7%)	6% (0·7%)
	2 to 5	32% (17%)	32% (9%)	25% (6%)	19% (4%)	16% (3%)
	6 to 10	13% (13%)	7% (9%)	20% (8%)	19% (6%)	16% (5%)
	more than 10	16% (61%)	31% (79%)	41% (85%)	49% (89%)	58% (92%)

example, the percentage of distinct L/R-allophones in the test set which occur more than 10 times in the training set is about twice of that of l/r-allophones, and the percentage is increased to about four-fold from lr-allophones to LR-allophones (compared to Table I). The total number of distinct L/R-allophones and LR-allophones in the test set is reduced from 612 to 194, and from 1500 to 468, respectively. It is clear that for most of the L/R-allophones and LR-allophones contained in test data, the corresponding HMMs will be reliable due to the high context coverage. This fact, combined with the higher accuracy of allophonic HMMs in capturing context-dependent acoustic variation of phonemes, would lead to improved recognition performance. This expectation has been confirmed by the experiments reported next.

5. Speech recognition experiments

5.1. Overview of the recognizer and speech data

The allophonic HMMs discussed in previous sections were evaluated in a speaker-dependent isolated-word recognizer with a 75 000-word English vocabulary. Overviews of the recognizer have been published previously (Deng *et al.*, 1988*a*; Gupta, Lennig & Mermelstein, 1988), and will be only briefly reviewed here. 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 (i.e. all words are considered *a priori* equally probable) or the trigram language model trained with a 57-million word text. The language model has a perplexity of about 900. The context-dependent allophonic HMMs described in this paper, as well as phonemic HMMs for comparison purposes, were used only at the exact likelihood scoring stage. Since the phonetic transcription of each candidate word generated by the fast search algorithm is known to the recognizer, so is the complete phonetic context of the constituent phones. This allows correct selection of one of the allophonic HMMs for each phoneme to score the unknown observation sequence.

Speech is recorded in a quiet sound booth using a Crown PZM microphone, low-pass filtered at 7 kHz, and sampled at 16 kHz. A Hamming window with a width of 25·6 ms is applied at 10 ms intervals. For each Hamming window, a 15-dimensional feature vector is computed. The vector consists of seven mel-frequency cepstral coefficients (Davis & Mermelstein, 1980), augmented by their differences over time and by the difference of loudness over time (Gupta, Lennig & Mermelstein, 1988).

Training and test data from three speakers, one male and two females, comprise natural-language sentences read from texts selected randomly from magazines, books and newspapers. The number of words (training and test data) recorded from each speaker vary between 2000 and 5000.

5.2. Recognition results

Table III shows the recognition accuracy obtained by the context-independent phonemic and the context-dependent allophonic HMM-based, speaker-dependent recognizers. Listed in the table are the recognition accuracies 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. Results are given for three speakers. For speaker CA, the training and the test data are those for the context-coverage analysis described in Section 4.

The benchmark system with phonemic HMMs has been described previously (Deng *et al.*, 1988*a*; Gupta, Lennig & Mermelstein, 1988). The L/R-allophonic and LR-allophonic HMMs have been described in the preceding sections of this paper. The results shown in Table III indicate that allophonic HMMs consistently outperform phonemic HMMs. The performance improvement is particularly noticeable when there are a large amount of training data (e.g. over 2500 words). In this case, averaged over speakers CA and AM, L/R-allophonic HMMs reduce recognition errors (compared with phonemic HMMs) by 18% for the use of uniform language model, and 26% for the use of the

TABLE III. Comparison of recognition rates of the context-dependent allophonic and context-independent phonemic HMM recognizers. Results are given for the use of uniform language model and trigram language model separately

Speaker (*test size*)	Training size	Phonemic HMMs		L/R-allophonic HMMs		LR-allophonic HMMs	
		uniform	trigram	uniform	trigram	uniform	trigram
CA (female)	717	67·9%	80·6%	70·0%	83·5%	69·7%	82·5%
(1090 words)	1532	70·2%	85·0%	75·0%	88·5%	78·1%	88·0%
	2347	70·6%	86·8%	75·5%	89·1%	80·9%	90·4%
	3098	70·8%	86·8%	76·0%	89·0%	82·7%	91·3%
	3880	70·7%	86·7%	76·1%	89·1%	83·0%	91·9%
AM (male)	1100	54·4%	78·0%	54·0%	78·0%	53·4%	77·0%
(698 words)	2039	68·2%	81·0%	73·0%	84·0%	72·0%	83·0%
	2742	68·7%	81·9%	74·2%	88·1%	76·1%	86·4%
MA (female) (586 words)	1600	79·0%	90·4%	83·1%	91·6%	83·8%	89·6%

trigram language model. LR-allophonic HMMs reduce the error rate further, by 35% and 33% 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 on the recognition accuracy. The context coverage is mainly determined by the training size (see Table II). The effects, demonstrated by the results in Table III, can be summarized as follows. An increase in the training size and consequently in the degree of context coverage tends to improve the recognition accuracy for both phonemic and allophonic HMMs. The improvement tends to saturate a certain training size. The saturation level is the lowest for phonemic HMMs, intermediate for L/R-allophonic HMMs, and the highest for LR-allophonic HMMs.

It is noted from Table III that in some cases L/R-allophonic HMMs produce higher recognition accuracy than LR-allophonic HMMs. This is unexpected since LR-allophonic HMMs in general capture more detailed coarticulatory information than do L/R-allophonic HMMs. We suspect that the poorer recognition accuracy for LR-allophonic HMMs results from unreliable estimation of mean vectors in certain LR-allophonic HMMs, 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 HMMs perform worse than L/R-allophonic HMMs 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 III). In the present course of study, we generally observed that model robustness has to be ensured to take full advantage of any powerful language model, while detailed and more accurate models (despite a lesser degree of robustness) manifest more strongly when no language models are used.

The interpolation method described by Equation (5) in Section 3 has been used to smooth the estimates of the mean vectors in training LR-allophonic HMMs. This experiment was carried out only for speaker AM. The results however, show that the recognition accuracy obtained by smoothing the mean vectors is about the same as without the smoothing.

6. Discussion and summary

For HMM-based large vocabulary speech recognition, representing phonemic units by HMMs is attractive for many reasons, the principal one being that the HMMs can be trained adequately with relatively small amounts of speech data. However, since phonemic classification is generally not consistent with acoustic distinction, the direct phonemic modeling approach has serious limitations. Context-dependent allophonic modeling is currently the most promising way in overcoming these limitations.

The context-dependent allophonic modeling approaches proposed so far can be classified into two main categories. The first, the structure-based approach, attempts to use innovative designs of the HMM structure according to the well-established acoustic-phonetic knowledge about speech. Use of such knowledge permits construction of more faithful representation of speech by HMMs, yet does not generally reduce the reliability of the representation. Although improvements in recognition accuracy for particular classes of phonemes have been made (Deng *et al.*, 1988*a*; Deng, Lennig & Mermelstein, 1989; Deng *et al.*, 1989 Deng, Lennig & Mermelstein, 1990) the ultimate large-scale success of this approach requires more thorough understanding of areas of phonetics

relevant to speech recognition and more skilful incorporation of the knowledge into the statistical modeling framework.

The second class of context-dependent allophonic modeling approaches, of which the present study belongs, is based on enumerating the phonetic contexts which are considered important in contributing to acoustic distinctions. Since the immediate neighboring phonemes tend to have the strongest coarticulatory effects on the phoneme under consideration, lr-allophones and l/r-allophones are often used. One obvious problem for this type of context-dependent modeling is that as more contexts are enumerated, the number of context-dependent models is increased as well, thus either making the models trained from a fixed amount of training data potentially less reliable, or practically reverting to less accurate context-independent models by any types of smoothing techniques.

To overcome this problem, we have developed two novel techniques in implementing context-dependent phonetic HMMs in a very large vocabulary speech recognizer. The first technique is to merge phonological contexts, guided by the acoustic-phonetic knowledge about the similarity in coarticulatory effects of different phonemes on their adjacent phonemes. We have presented a detailed analysis on the overlap of common contexts in the training and in the test data. The results of this analysis provide a quantitative assessment on the improved reliability of the allophonic HMMs achieved by context merging.

The second technique to improve reliability of context-dependent allophonic HMMs is to smooth the covariance matrices by tying them in Gaussian HMMs across all allophones. We show that after the tying, additional smoothing on the mean vectors in allophonic HMMs by linearly interpolating them with those in context-independent HMMs becomes unnecessary. The possibility of smoothing HMM output distributions by tying covariance matrices is an important advantage of the Gaussian HMM over the HMM with non-parametric discrete output distributions. In the latter case, the only viable smoothing method is the interpolation between the context-dependent and context-independent HMM parameters (Schwartz *et al.*, 1984). In our previous experiments (unpublished), we carried this interpolation method through all parameters in Gaussian HMMs, including the covariance matrices. We achieved no error reduction using the context-dependent HMMs obtained in this way. Therefore, one of the main contributions of this study is to demonstrate the effectiveness of tying covariance matrices across contexts in constructing context-dependent Gaussian HMMs.

Significant improvement in recognition accuracy has been achieved using allophonic HMMs constructed by the above two techniques. The error rate reduction is 26% and 33% for the use of L/R-allophonic and LR-allophonic HMMs, respectively, when the training data consist of over 2500 words. This size of training data manifests over 90% and 70% overlap (occurrence frequencies more than five) of L/R-allophonic and LR-allophonic contexts with the test data (calculated from the statistics shown in Table II). For smaller amounts of training data and hence lower context overlap, the error reductions become progressively smaller.

Further improvements in recognition accuracy can be expected beyond the work presented here. First, our method for merging contexts is based on intuition grounded in phonetic knowledge, and may be inadequate in terms of optimizing speech recognition performance. It is desirable to develop context-merging procedures which are aimed directly at improving phonetic distinction. The automatic clustering algorithm devised by Lee (1988) for a medium-sized vocabulary speech recognizer is not applicable to very

large vocabulary speech recognition since the algorithm requires each allophone to be present in the training data. In our vocabulary containing 86 000 words and 92 000 phonemic transcriptions, more than 17 000 triphones have been counted. To apply the automatic algorithm for clustering these triphones in a speaker-dependent system, we would probably need speech material from the amount of natural text equivalent to a full-length novel. Second, due to the merge of contexts, the acoustic variability within each merged context would necessarily be greater than that without the merge. This suggests that the unimodal Gaussian assumption for the output distributions in context-dependent HMMs may be inappropriate. Use of Gaussian mixtures for the output distributions may result in more accurate HMMs. Finally, we intend to combine the structure-based and the enumeration-based context-dependent modeling approaches in order to achieve a better compromise between the model accuracy and the model reliability.

References

Bahl, L. R., Bakis, R., Cohen, P. S., Cole, A. G., Jelinek, F., Lewis, B. L. & Mercer, R. L. (1980). Further results on the recognition of a continuously read natural corpus. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 872–875.

Bahl, L. R., Jelinek, F. & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-5**, 179–190.

Bahl, L. R., Brown, P. F., de Souza, P. V., Picheny, M. A. & Mercer, R. L. (1988). Acoustic Markov models used in the Tangora speech recognition system. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 497–500.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, **3**, 1–8.

Baum, L. E. & Sell, G. R. (1968). Growth transformations for functions on manifolds. *Pacific Journal of Mathematics*, **27**, 211–227.

Chow, Y. L., Dunham, M. D., Kimball, O. A., Krasner M. A., Kubala, G. F., Makhoul, J., Price, P. J., Roucos, S. & Schwartz, R. M. BYBLOS: The BBN continuous speech recognition system. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 89–92.

Davis, S. B. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-28**, 357–365.

Deng, L., Kenny, P., Lennig, M., Gupta, V. & Mermelstein, P. (1988*a*). 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.

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

Deng, L., Kenny, P., Lennig, M., Gupta, V. & Mermelstein, P. (1989). A locus model of coarticulation in an HMM speech recognizer. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 97–100.

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

Deng, L., Lennig, M. & Mermelstein, P. (1990). Modeling microsegments of stop consonants in an HMM-based speech recognizer. *Journal of the Acoustical Society of America*. In press.

Derouault, A. (1987). Context-dependent allophonic Markov models for large vocabulary speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 360–363.

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

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, **64**, 532–556.

Lee, C. H., Soong, F. K. & Juang, B. H. (1988). A segment model approach to speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 501–504.

Lee, K. F. (1988). Large vocabulary speaker-independent continuous speech recognition: The SPHINX system. Ph.D. dissertation, Computer Science Department, Carnegie Mellon University.

Lee, K. F. & Hon, H. W. (1988). Large vocabulary speaker-independent continuous speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 123–126.

Levinson, S. E., Rabiner, L. R. & Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell System Technical Journal*, **62**, 1035–1074.

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

Lippmann, R. P., Martin, E. A. & Paul, D. P. (1987). Multi-style training for robust isolated-word speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 705–708.

Merialdo, B. (1987). Speech recognition with very large size dictionary. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 364–367.

Murveit, H. & Weintraub, M. (1988). Speaker-independent connected-speech recognition using hidden Markov models. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 115–118.

Paul, D. B. & Martin, E. A. (1988). Speaker stress-resistant continuous speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 283–286.

Rabiner, L. R., Wilpon, J. G. & Soong, F. K. (1988). High performance connected digit recognition using hidden Markov models. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 122–125.

Rabiner, L. R., Juang, B. H., Levinson, S. E. & Sondhi, M. M. (1985). Some properties of continuous hidden Markov model representations. *Bell System Technical Journal*, **64**, 1251–1270.

Schwartz, R. M., Chow, Y. L., Roucos, S., Krasner, M. & Makhoul, J. (1984). Improved hidden Markov modeling of phonemes for continuous speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 35.6.1–35.6.4.