IMPROVED ACOUSTIC MODELING FOR CONTINUOUS SPEECH RECOGNITION

C.-H. Lee, E. Giachin† *, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg*

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974

ABSTRACT

We report on some recent improvements to an HMM-based, continuous speech recognition system which is being developed at AT&T Bell Laboratories. These advances, which include the incorporation of inter-word, context-dependent units and an improved feature analysis, lead to a recognition system which achieves better than 95% word accuracy for speaker independent recognition of the 1000-word, DARPA resource management task using the standard word-pair grammar (with a perplexity of about 60). It will be shown that the incorporation of inter-word units into training results in better acoustic models of word juncture coarticulation and gives a 20% reduction in error rate. The effect of an improved set of spectral and log energy features is to further reduce word error rate by about 30%. We also found that the spectral vectors, corresponding to the same speech unit, behave differently statistically, depending on whether they are at word boundaries or within a word. The results suggest that intra-word and inter-word units should be modeled independently, even when they appear in the same context. Using a set of sub-word units which included variants for intra-word and inter-word, context-dependent phones, an additional decrease of about 10% in word error rate resulted.

1. INTRODUCTION

In the past few years there have been proposed a number of systems for large vocabulary, speaker-independent, continuous speech recognition which have achieved high word recognition accuracy [1-5]. The approach to large vocabulary speech recognition we adopt in this paper is a pattern recognition based approach. The basic speech units in the system use phonetic labels and are modeled acoustically based on a lexical description of words in the vocabulary. No assumption is made, *a priori*, about the mapping between acoustic measurements and sub-word linguistic units such as phonemes; such a mapping is entirely learned via a finite training set of utterances. The resulting speech units, which we call *phone-like units* (PLU's) are essentially acoustic descriptions of linguistically-based units *as represented in the words occurring in the given training set.*

In the baseline system reported in [1], acoustic modeling techniques for intra-word context-dependent PLU's were discussed. The focus of this paper is to extend the basic acoustic modeling techniques developed in [1] to include modeling of word juncture coarticulation and to incorporate higher-order time derivatives of cepstral and log energy parameters into the feature vector in order to improve speech recognition performance.

We tested the improved acoustic modeling techniques on speaker-independent recognition of the DARPA Naval Resource Management task using both the word-pair (WP) and the no grammar (NG) conditions. For the FEB89 test set using the WP grammar, the word accuracy improved from 91.3% to 95.0% when both the inter-word context-dependent PLU's and an improved feature analysis were incorporated into the baseline system. We also observed that, for the first time, over 70% sentence accuracy was achieved. The same level of improvement was also obtained for the OCT89 and the JUN90 test sets.

2. BASELINE RECOGNITION SYSTEM

There are three main modules in the baseline recognition system, namely a feature analysis module, a word-level acoustic match module and a sentence-level language match module. The speech is filtered from 100 Hz to 3.8 kHz, and sampled at an 8 kHz rate and converted into a sequence of feature vectors at a frame rate of 10 msec. Each (24-element) feature vector consists of 12 liftered cepstral coefficients and 12 delta cepstral coefficients. The reader is referred to [1] for a detailed description of the acoustic analysis procedure. We will show later that higher order time derivatives of cepstral

† Now with CSELT, Torino, Italy.

and energy parameters can also be incorporated into the feature vectors to improve recognition performance.

The word-level match module evaluates the similarity between the input feature vector sequence and a set of acoustic word models to determine what words were most likely spoken. The word models are generated via a lexicon and a set of sub-word models. In our current implementation, we use a slightly modified version of a lexicon provided by CMU. Every word in the vocabulary is represented by exactly one entry in the lexicon, and each lexical entry is characterized by a linear sequence of phone units. Each word model is composed as a concatenation of the sequence of sub-word models according to its corresponding lexical representation. A set of 47 context-independent phone labels is extracted from the lexicon and is used as the set of basic speech units throughout this study. Context-dependent units are also derived from this basic unit set.

The sentence-level match module uses a language model (based on a set of syntactic and semantic rules) to determine the word sequence in a sentence. In our current implementation, we assume that the language model is fixed and can be represented by a finite state network (FSN). The word-level match module and the sentence-level match module work together to produce the mostly likely recognized sentence.

2.1 PLU Modeling

Each PLU is modeled as a left-to-right hidden Markov model (HMM). All PLU models have three states, except for the silence PLU model which has only one state. Furthermore, no transition probabilities are used; forward and self transitions from a state are assumed equally likely. The state observation density of every PLU is represented by a multivariate Gaussian mixture density with a diagonal covariance matrix.

To avoid singularity caused by an under-estimation of the variance, we replaced all estimates whose values were in the lowest 20% with an estimate of the 20 percentile value of the variance histogram. It can be shown [7] that such a variance clipping strategy results in the maximum a posteriori estimate for the variance parameter, if we assume that the variance is known a priori to exceed a fixed threshold (which has to be estimated from a set of training data).

Training of the set of sub-word units is accomplished by a modified version of the segmental *k*-means training procedure [1,8]. Since acoustic segments for different PLU labels are assumed independent, models for each unit can be constructed independently. By doing so, the computation requirement for HMM training is much less than the standard forward-backward HMM training.

2.2 Creation of Context Dependent PLU's

The idea behind creating context dependent PLU's is to capture the local acoustic variability associated with a known context and thereby reduce the acoustic variability of the set of PLU's. One of the earliest attempts at exploiting context dependent PLU's was in the BBN BYBLOS system where left and right context PLU's were introduced [9]. Given the set of 47 PLU's, a total of $47^3 = 103823$ CD PLU's could be obtained in theory. In practice, only a small fraction of them appear in words for a given task. However the number of CD PLU's is still very large (on the order of 2000-7000), making even a reasonable amount of training material insufficient to estimate all the CD PLU models with acceptable accuracy. Several techniques have been devised to overcome these training difficulties. Perhaps the simplest way is to use a unit reduction rule [1] based on the number of occurrences of a particular unit in the training set.

2.3 Experimental Setup and Baseline Results

Throughout this study, we used the training and testing materials for the DARPA naval resource management task. The speech database was originally provided by DARPA at a 16 kHz sampling rate. To make the database compatible with telephone bandwidth, we filtered and down-sampled the speech to an 8 kHz rate before analysis. The training set consists of a set of 3990 read sentences from 109 talkers (30-40 sentences/talker). For most of our evaluations, we used the so-called FEB89 test data which consisted of 300 sentences from 10 new talkers (30 sentences/talker) as distributed by DARPA in February 1989.

A detailed performance summary of the baseline system and the techniques we used to achieve such a performance is given in [1]. In order to provide a consistent base of performance for comparison with the new results, we used a threshold of 30 in the unit reduction rule for all performance evaluations. When no inter-word units were used, a threshold of 30 resulted in a set of 1090 PLU's. With this set of units, we obtained a word accuracy of 91.3% and a sentence accuracy of 58.7% on the FEB89 test set.

3. WORD JUNCTURE MODELING

Coarticulation phenomena arising at the boundary between words are a major cause of acoustic variability for the initial and final parts of a word when spoken in continuous speech. If this type of contextual variability is not adequately represented in the recognition system, errors are likely to occur. This is the case, for instance, when words in the dictionary are phonetically transcribed according to their pronunciations in isolation. The solution to this problem is to provide a more precise phonetic representation of the region across word boundaries. Different approaches may be taken to achieve this goal, depending on which type of coarticulation phenomenon is handled. In [10] we characterized two different types of pronunciation changes at word junctures, namely *soft* and *hard* changes. In soft changes, the alteration of a phone due to neighboring phones is comparatively small and the actual realization is perceived as a variation of the original phone rather than a transformation to a different phone. This alteration is of the same nature as the one observed in intra-word phones. By contrast, in hard changes, a boundary phone may undergo a complete deletion or a substitution by a totally different phone. It was experimentally shown that phonological rules were effective in coping with errors generated by hard changes (a 10% error rate reduction [10]). However for soft changes, it has been shown that the use of *context-dependent phone-like units* is most effective. Such units specify context-independent phones according to their context, i.e. the preceding phone and the following phone. Context-dependent phones are not very effective with hard changes because these changes are comparatively rare and the training material is not sufficient to model the units that would represent them. Soft changes occur more frequently in the training set and hence we can model such changes similar to the way we model the set of intra-word units.

3.1 Word Juncture Phonological Rules

In the current set-up, phonological rules are employed with the set of 47 context-independent phone units. The phonological rules have been implemented both in the training as well as in the recognition procedure. The reason for introducing the rules during training is that they give a more precise phonetic transcription of the training sentences and hence allow a better segmentation and, consequently, better estimation of model parameters. In preliminary experiments, it was seen that about 50% of the rule corrective power is actually due to training.

Recognition is based on a modified version of the frame-synchronous beam search algorithm described in [1]. A detailed summary of the experimental results using phonological rules can be found in [10]. Based on the set of 47 context-independent PLU's, training on 3200 utterances and testing on 750 utterances, we found that a 10% error rate reduction can be obtained. Although the set of phonological rules was applied rarely in both training and in recognition, the application of phonological rules was shown very effective in dealing with the hard word juncture changes which are difficult to model statistically.

3.2 Inter-word Context-Dependent Units

Several schemes have been proposed for modeling and incorporating inter-word, context-dependent phones into a continuous speech recognition system. The methodology we adopted for inter-word unit modeling is described in detail in [11]. Since the initial phone of each word depends on the final phone of the preceding word and similarly for the final phone of each word, there are many more context-dependent, inter-word units than intra-word units. In the 109 speaker training set, there are roughly 5,500 inter-word units compared to only 1,800 intra-word units. Therefore the training issues outlined above become even more complicated. Moreover, in recognition, since it is not known *a priori* which words will precede or follow any given word, the words have to be represented with all their possible initial and final phones. In addition, for reasons that will be explained below, words consisting of one single phone (e.g. "a") have to be treated with special care; thus further increasing the complication of the recognition and training procedures.

The complete set of units, including both intra-word and inter-word units, was chosen according to the same unit reduction rule described in [1]. By varying the value of the threshold T different PLU inventories can be obtained and tested. Based on the reduction rules, the inter-word units generated include double-context units (triphones), single-context units (diphones) and context-independent units (monophones). A typical implementation discussed in the literature [2] use only triphones and monophones; however we found diphones quite helpful in our implementation. Using the same threshold, i.e. T=30, we obtain a set of 1282 units, including 1101 double-context units, 99 left-context units, 35 right-context units and 47 context-independent units. It is noted here that in our modeling strategy, every acoustic segment in the training data has only one PLU label and is used only once in training to create a sub-word model of the particular phone unit.

Training is based on a modified version of the segmental k-means training procedure [1,8]. However, due to the presence of inter-word units, cross-word connections need to be handled properly in the segmentation part of the procedure. Segmentation is carried out on the finite state network (FSN) that represents each utterance in terms of PLU's. Since context-dependent inter-word PLU's are available, there is no discontinuity between words. However, optional silences are allowed between words; in such cases PLU's having either a left or right silence context are used.

The recognizer used in our research is based on a modified version of the frame-synchronous beam search algorithm [1]. However due to the presence of the complicated inter-word connection structure, the finite state network representing the task grammar is converted into an efficient *compiled network* to minimize computation and memory overhead. The reader is referred to [12] for a detailed description of the recognizer implementation. The recognizer is run with no grammar (i.e. every word can follow every other word) or with a word pair grammar.

A detailed summary of the experimental results using inter-word units can be found in [11]. Based on the set of 1282 PLU's, the recognition performance was 93.0% word accuracy and 63.7% sentence accuracy for the FEB89 test set. A comparison with the results obtained without using inter-word units (i.e. the 1090 PLU set) shows that a 20% error rate reduction resulted.

After a close examination of the recognition results, several observations can be made. First, there are less errors on function words (e.g. "a", "the", etc.), because a better word juncture coarticulation model gives a better representation of those highly variable short words. Second, we observed a much higher likelihood score when inter-word units are used in both training and recognition. We also found that, when an utterance was misrecognized, the likelihood difference between the recognized and the correct strings was smaller than that with only intra-word units. This shows that the misrecognized utterance is more likely to be corrected when better acoustic modeling techniques are incorporated into the system. We now discuss some techniques for obtaining an improved feature set.

4. IMPROVED FEATURE ANALYSIS

So far, we have discussed the criteria for the selection of the set of fundamental units, shown how to expand the set to include both intra-word and inter-word, context-dependent PLU's and discussed how to properly model these units. In this section, we focus our discussion on an improved front-end feature analysis. Since we are using a continuous density HMM approach for characterizing each of the sub-word units, it is fairly straightforward to incorporate new features into our feature vectors. Specifically, we study the incorporation of higher order time derivatives of short-time cepstral features and log energy features, such as the second cepstral derivatives *(delta-delta cepstrum)*, the log energy derivative *(delta energy)*, and the second log energy derivative *(delta-delta energy)*, into our system.

Second Order Cepstral Time Derivatives

The incorporation of first order time derivatives of cepstral coefficients has been shown useful for both speech recognition and speaker verification. Thus we were interested in investigating the effects of incorporating higher order cepstral time derivatives. There are several ways to incorporate the second order time-derivative of the cepstral coefficients. All the existing approaches evaluate the second derivatives (called *delta-delta cepstrum*) as the time derivatives of the first order time derivative - so called delta cepstrum. The degree of success in using such a strategy for the delta-delta cepstrum computation was mixed. The only continuous speech recognition system which used using the delta-delta cepstral features was reported by Ney [13]. Ney tested the system using speaker independent recognition of the DARPA naval resource management task, and showed a very significant improvement when testing the recognizer without using any grammar. However, for the word-pair grammar, there was no significant improvement in performance.

In our evaluation, we used a 3-frame window (i.e. an overall window of 70 msecs for both delta and delta-delta cepstrum). The m^{th} delta-delta cepstral coefficient at frame l was approximated as

$$\Delta_2\hat{c}_l(m) = K \cdot \left[\Delta\hat{c}_{l+1}(m) - \Delta\hat{c}_{l-1}(m)\right] \qquad (1)$$

where $\Delta\hat{c}_l(m)$ is the estimated m^{th} delta cepstral coefficient evaluated at frame l, and K is a scaling constant which was fixed to be 0.375 (no optimization was attempted to find a better value of the normalization constant to optimize the k-means clustering part of the training algorithm).

We augment the original 24-dimensional feature vector with 12 additional delta-delat cesptral features giving a 36-dimensional feature vector. We then tested this new feature analysis procedure on the resource management task using the word-pair grammar. We tested cases both with and without the use of inter-word units. The per speaker word accuracies are summaried in Table 1. The effect of adding the delta-delta cepstral parameters on recognition performance (using both inter-word and intra-word units) on the set of 1282 PLU's, was that the performance improved from 93.0% (column 2) to 93.8% (column 3) on the FEB89 test set. The error rate reduction for the latter case was not as much as the former case. For the inter-word case (1282 PLU), we ran four iterations of the segmental k-means training procedure. We also tested models generated in all four iterations. It is worth noting that the best average performance was achieved using the model generated from the fourth iteration (shown in column 3). When comparing performance on a per speaker basis, the results showed that the addition of the delta-delta cepstral features to the overall spectral feature vector is not always beneficial for each speaker. For example, there is a significant performance degradation for speaker 1 (cmh18) in the 1282 PLU case. We also observed a large variability in speaker performance using models obtained from various iterations. We therefore manually extracted the best performance among all four iterations, for each speaker, and list the results in column 4 of Table 1. It is interesting to note that the average best performance is much better (16% reduction in error rate) than the results listed in column 3. Our conjecture is that the second order cepstral analysis produces very noisy observations based on a 30 msec window and a 10 msec frame shift. Another concern is the effectiveness of each of the additional features. In Ney [13], a preselected set of delta and delta-delta cepstral features was used. To be more effective, an automatic feature selection algorithm should be used to determine the relative importance of all spectral analysis features.

	1282 PLU set		
Speaker ID	DCEP	DDCEP	BEST
cmh18	90.9	88.7	90.9
dml01	89.7	92.6	94.5
dwa05	89.6	90.4	90.4
esg04	96.7	97.9	99.2
gaw07	95.1	96.7	96.7
gmb05	93.5	95.1	95.9
hlm02	96.3	96.6	97.6
jdh06	92.7	93.9	95.1
kls01	93.4	93.4	95.5
lns03	91.9	91.9	91.9
Average	93.0	93.8	94.8

Table 1. Delta-delta cepstrum test results

A few observations can be made from the above results. Overall it can be seen that the incorporation of second order time derivatives of cepstral parameters improves recognition performance significantly. However the second order time derivatives are very noisy in the sense that they produce features which are not always beneficial and they are not stable over the segmental k-means training iterations. From the last column of Table 1, it is clear that improved performance could be achieved if we could stablize the features across training iterations. One way to improve the feature analysis is to carefully select new features so that only features that are useful for discrimination are included in the feature vector. Another way is to combine features through a principal component analysis.

4.1 Log Energy Time Derivatives

The first order time derivatives of the log energy values, known as delta energy, have been shown useful in a number of recognition systems. Most systems use both energy and delta energy parameters as features. In order to use the energy parameter, careful normalization is required. In our baseline system, the energy parameter was normalized syllabically. We did not include the energy parameter in the feature vector; instead we used the energy parameter to assign a penalty term to the likelihood of the observed feature vector. However, we have found that the delta and delta-delta energy parameters are more robust and more effective recognition features. Similar to the evaluation of the delta cepstrum, the delta energy at frame l is approximated as a linear combination of the energy parameters in a 5 frame window centered at frame l. Since the energy parameter has a wider dynamic range in value, we used a smaller constant (0.0375) for the evaluation of the delta energy. Again, we did not attempt to optimize the k-means clustering part by adjusting the normalization constant.

The second order time derivatives of the energy parameters, called the delta-delta energy, are computed similar to the way the delta-delta cepstral features are evaluated. Starting with the 24-element feature vector, by adding delta-delta cepstrum, delta energy and delta-delta energy to the feature set, for every frame l, we have a 38-element feature vector O_l of the form

$$O_l = \{\hat{c}_l(1{:}12),\ \Delta\hat{c}_l(1{:}12),\ \Delta_2\hat{c}_l(1{:}12),\ \Delta e_l,\ \Delta_2 e_l\} \quad (2)$$

where $\hat{c}_l(1{:}12)$ and $\Delta\hat{c}_l(1{:}12)$ are the 12 liftered cepstral coefficients and the 12 delta cepstral coefficients; and $\Delta_2\hat{c}_l(1{:}12)$, Δe_l and $\Delta_2 e_l$ are the additional 12 delta-delta cepstral coefficients, the delta energy parameter and the delta-delta energy parameter at frame l respectively.

We tested the use of energy time derivatives on the FEB89 test with the word pair grammar. All the tests used the set of 1282 PLU's, and the test results are summarized in Table 2. When compared with the results shown in Table 1, we observed a very significant overall improvement when delta energy is incorporated (shown in column 2). We also note that the improvement varies from speaker to speaker. For example, delta energy helps eliminate a lot of word errors for speaker 1 (cmh18). When delta-delta energy is added to form a 38-dimensional feature vector, we observe the same effect as shown in Table 1, i.e. delta-delta energy is not always beneficial. However, for some talkers (cmh18, dwa05 and lns03) the improvement is significant. There is no overall improvement; however all the test speakers achieve over 92% word accuracy. We show in column 4, the best achievable performance for each talker using various feature combinations. The best achievable average performance is over 96% word accuracy.

Speaker ID	+DENG	+DDENG	BEST
cmh18	93.0	95.2	95.7
dml01	95.5	94.5	95.5
dwa05	91.5	92.2	93.0
esg04	97.9	97.5	99.2
gaw07	95.9	96.3	97.1
gmb05	96.7	95.9	96.7
hlm02	96.6	95.9	97.6
jdh06	94.7	94.3	95.5
kls01	96.3	94.7	97.1
lns03	91.5	92.6	93.0
Average	94.9	94.9	96.1

Table 2. Log energy time derivatives test results

5. POSITION-DEPENDENT UNIT MODELING

For all our experiments we have selected the set of basic speech units based on context. However, it is believed that the spectral features of units within words behave differently acoustically from those of units at the word boundaries even when the units appear in the same context. We investigated this conjecture by selecting intra-word and inter-word units independently based on the same unit reduction rule. We call such a selection strategy *position-dependent unit selection*. With a threshold of 30, we obtained a total of 1769 units, including 913 intra-word and 856 inter-word units. For the intra-word units, we have 639 units in double context, 98 left-context and 176 right-context units and 47 context-independent units. For the inter-word units, we end up with 480 units in double context, 310 left-context, 2 right-context units and 46 context-independent units. We use the same modeling strategy described in Section 3, except that when creating the FSN for segmentation and recognition, only inter-word units can appear at the word boundaries and only intra-word units can appear within words.

Using such a position-dependent unit selection strategy, we found that all the sub-word unit models are more focused in the sense that the spectral variability is less than the case in which we combine common intra-word and inter-word unit models. Two interesting observations are worth mentioning. First, when recognition is performed with the beam search algorithm, the number of alive nodes is much less in the 1769 PLU case than that in the 1282 PLU case. Second, the unit separation (in terms of likelihood) distance is larger for the 1769 PLU set than that for the 1282 PLU set. The unit separation distance is measured as follows. We first collect all sets of PLU's such that all the units which have the same middle phone symbol p are grouped into the same set S_p regardless of their context and environment. We compute the distance between each pair of units in a set. The distance between units P_2 and P_1 is defined as the difference between average likelihoods of observing all acoustic segments with label P_2 and observing all acoustic segments with label P_1 given the model for unit P_1. For each unit, we then define the *unit separation distance* as the smallest distance among all other units in the same set. When examing the histogram for the unit separation distances for the 1769 PLU set, we found a quite remarkable unit separation. Almost all the unit separation distances are larger than 2 (in log likelihood). It is also interesting to note that units appearing in the same context but in a different environment (i.e. intra-word versus inter-word) show the same behavior as units appearing in different context. The average unit separation distance is about 9. For the 1282 PLU set, the histogram plot is skewed to the left and the unit separation characteristics are not as pronounced as those of the 1769 PLU set.

Results on the resource management task using the 1769 PLU set showed a significant improvement (about 10% error reduction) in performance over the 1282 PLU set using both the WP grammar and the NG grammar. We also tested the OCT89 set which consists of 300 test utterance (30 each from a group of 10 speakers), and the JUN90 set which consists of 480 test utterances (120 each from a group of two female and two male talkers). Detailed results for all three test sets using both the WP and NG cases are summarized in Tables 3 and 4 respectively. A careful examination of the word error patterns shows that function words still account for about 60% of the word errors. The second dominant category, which accounts for more than 20% of the errors, are confusions involving the same root word (e.g. location versus locations, six versus sixth, Flint versus Flint's, chop versus chopped, etc.) appearing in different forms. This type of errors can easily be corrected with a simple set of syntactic and semantic rules.

Testing Set	FEB89	OCT89	JUN90
Word Corr.	96.1	96.3	95.6
Sub. Error	2.9	2.7	3.3
Del. Error	1.0	0.9	1.1
Ins. Error	0.7	0.9	0.5
Word Error	4.6	4.5	4.9
Sent. Acc.	74.7	73.7	71.5
Word Acc.	95.4	95.5	95.1

Table 3. WP test summary using the 1769 PLU set

Testing Set	FEB89	OCT89	JUN90
Word Corr.	81.8	81.9	78.9
Sub. Error	14.3	14.5	15.6
Del. Error	3.8	3.8	5.5
Ins. Error	2.1	2.6	1.4
Word Error	20.3	20.7	22.5
Sent. Acc.	27.3	26.3	24.2
Word Acc.	79.7	79.3	77.5

Table 4. NG test summary using the 1769 PLU set

It is interesting to note that for the speaker-independent part of the evaluation data, we achieved virtually the same level of performance, 95.5% word accuracy and over 70% sentence accuracy, for both the FEB89 and OCT89 test data. However, for the JUN90 test set, the word accuracy fell to 94.9%. The sentence accuracy of 70.2% was also considerably lower than that for the FEB89 and OCT89 test sets. This was due to the fact that one of the female test talkers (jrm08) gave a very poor result, namely 90.8% word accuracy. The median word accuracy among all four JUN90 test talkers was over 95.5%. Another possible explanation for this performance degradation is that the speaker-dependent part of the test utterances was generated from a set of sentence patterns whose context was significantly different from the set of sentences used to generate the 109-speaker training set. This brings up the issue of the adequacy of training data. The above discussion suggests that a task-specific training set is more useful than a task-independent training set. This agrees somewhat with results reported in [14], where the so called "vocabulary-independent" training procedure is effective only when most of the task-specific triphone contexts appear in the training corpus.

Even though only a small improvement in performance was obtained in our test, we believe the real benefit of incorporating position-dependent PLU modeling lies in the area of model prediction for units appearing rather infrequently in the training data. We are now experimenting with a *unit expansion rule*, which uses the unit reduction rule first (based on a lower count threshold) to get a set of *auxiliary speech units* which are the units to be expanded into a complete set. The only remaining issue is how to predict the model for units in the auxiliary set. For each existing model, we compute the likelihood of observing all the acoustic segments corresponding to each auxiliary unit. Then we assign the model that is the closest to each of the auxiliary speech units. The effectiveness of such an approach is yet to be evaluated.

6. SUMMARY

We have reported on several improvements to one of the speaker-independent, continuous speech recognition systems developed at AT&T Bell Laboratories. The improved acoustic modeling, including incorporation of inter-word units and an improved feature analysis, provided high word accuracies for all three DARPA evaluation sets using the word pair grammar. We have also developed a unit selection rule for selecting intra-word and inter-word units independently. We anticipate that with the proposed unit expansion rule, an even better set of units can be obtained which will further improve acoustic modeling techniques for continuous speech recognition.

Based on current recognition performance it seems fair to say that when task-specific training data are provided for acoustic modeling of the set of basic speech units using HMM's, high performance can be achieved for a large vocabulary, speaker-independent, continuous speech recognition task with a perplexity of about 60. However, there are still some open issues that need to be addressed. The reader is referred to a recent paper [15] for a discussion of some of those issues related to using HMM's for speech recognition. We list, in the following, a number of acoustic modeling issues which we believe to be essential for expanding the capabilities of our current continuous speech recognition system. They are: (1) Speech unit selection and modeling for task-independent applications; (2) Improved word discrimination based on some form of corrective training for continuous density HMM parameters (e.g. [16]); (3) Lexical modeling to deal with lexical variability in baseform pronunciation; and (4) Improved feature selection so that only those features useful to discrimination are included in the feature vector.

Our tasks so far have been mainly focused on speech recognition. We observed that short function words (e.g. "a", "the") are a major source of recognition errors. However, most of those errors can be corrected using a set of simple syntactic and semantic rules operating in a post-processing mode. For example, for the resource management task, we have developed a language decoder (decoupled from the acoustic decoder) that incorporates a set of simple rules. Our preliminary results [17] indicate that sentence accuracy for the FEB89 test set improved from 70% to close to 90% (with 98% word accuracy) when the top candidate string decoded using the word-pair grammar is used as input to this language decoder. When no grammatical constraints were used in speech decoding, the sentence accuracy improved from 24% to 67% (with 90% word accuracy). Except in cases where some key content words were misrecognized, the simple language analyzer properly decoded the noisy strings provided by the speech decoder without going back to the acoustic domain to request mismatch information. We believe the language decoder can be more effective if the speech decoder can provide more word and string hypotheses. One way to get more information is to use the *N*-best string search strategies. Another way is to construct, in acoustic decoding, a phone lattice and a word lattice that contain more word hypotheses, and then generate recognized strings according to the language constraints. The effectiveness of such approaches in real spoken language tasks, such as the DARPA Air Travel Information System (ATIS) task, is yet to be evaluated.

REFERENCES

1. C.-H. Lee, L. R. Rabiner, R. Pieraccini, J. G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition", *Computer Speech and Language*, **4**, pp. 127-165, 1990.

2. K. F. Lee, *Automatic Speech Recognition – The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, 1989.

3. D. B. Paul, ''The Lincoln Robust Continuous Speech Recognizer,'' *Proc. ICASSP-89*, Glasgow, Scotland, pp. 449-452, May 1989.

4. M. Weintraub *et al.*, ''Linguistic Constraints in Hidden Markov Model Based Speech Recognition,'' *Proc. ICASSP-89*, Glasgow, Scotland, pp. 699-702, May 1989.

5. V. Zue, J. Glass, M. Phillips, and S. Seneff, ''The MIT Summit Speech Recognition System: A Progress Report,'' *Proc. Speech and Natural Language Workshop*, pp. 179-189, Feb. 1989.

6. L. R. Rabiner, ''A Tutorial on Hidden Markov Models, and Selected Applications in Speech Recognition,'' *Proc. IEEE*, Vol. 77, No. 2, pp. 257-286, Feb. 1989.

7. C.-H. Lee, C.-H. Lin and B.-H. Juang, "A Study on Speaker Adaptation of the Parameters for Continuous Density Hidden Markov Models", to appear in *IEEE Trans. on Acoustic, Speech and Signal Proc.*

8. L. R. Rabiner, J. G. Wilpon, and B. H. Juang, ''A Segmental *K*-Means Training Procedure for Connected Word Recognition,'' *AT&T Tech. J.*, Vol. 65, No. 3, pp. 21-31, May-June 1986.

9. R. Schwartz *et al.*, ''Context Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech,'' *Proc. ICASSP 85*, Tampa, Florida, pp. 1205-1208, March 1985.

10. E. Giachin, A. E. Rosenberg and C.-H. Lee, "Word Juncture Coarticulation Modeling Using Phonological Rules for HMM-based Continuous Speech Recognition", *Proc. ICASSP 90*, pp. 737-740, Albuquerque, NM, April 1990.

11. E. Giachin, C.-H. Lee, L. R. Rabiner and R. Pieraccini, "Word Juncture Modeling Using Inter-Word Context-Dependent Phone-Like Units", submitted for publication.

12. R. Pieraccini, C.-H. Lee, E. Giachin and L. R. Rabiner, "Implementation Aspects of Large Vocabulary Recognition Based on Intra-word and Inter-word Phonetic Units", *Proc. DARPA Speech and Natural Language Workshop*, Somerset, PA, June 1990.

13. H. Ney, ''Acoustic-Phonetic Modeling Using Continuous Mixture Densities for the 991-Word DARPA Speech Recognition Task,'' *Proc. ICASSP 90*, pp. 713-716, Albuquerque, NM, April 1990.

14. H. W. Hon, K. F. Lee, and R. Weide, ''Towards Speech Recognition Without Vocabulary Specific Training,'' *Proc. EuroSpeech 89*, pp. 481-484, Paris, France, September 1989.

15. B.-H. Juang and L. R. Rabiner, ''Issues in Using Hidden Markov Models for Speech Recognition,'' to appear in *Advances in Speech Signal Processing*, S. Furui and M. Sondhi editors, Marcel Dekker Inc., New York, 1990.

16. S. Katagiri and C.-H. Lee ''A New HMM/LVQ Hybrid Algorithm for Speech Recognition,'' to appear in *Proc. GLOBECOM-90*, San Diego, CA, December 1990.

17. R. Pieraccini, K.-Y. Su and C.-H. Lee, unpublished work.