# IMPROVED ACOUSTIC MODELING FOR SPEAKER INDEPENDENT LARGE VOCABULARY 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, USA

## ABSTRACT

We report on some recent improvements to an HMM-based, continuous speech recognition system developed at AT&T Bell Laboratories. These advances, which include the incorporation of inter-word context-dependent units and position-dependent units and an improved feature analysis, lead to a recognition system which gives a 95% word accuracy and 75% sentence 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). With the improved acoustic modeling of sub-word units, the overall error rate reduction was over 42% compared with the performance results reported in our baseline system [1]. The best results we obtained so far, using the word pair grammar, gave 95.2% average word accuracy for the three DARPA evaluation sets [2]. The same improved acoustic modeling techniques was also found effective for small vocabulary tasks. For the TI connected digit test, we achieved a 50% string error reduction over out best result when the improved feature analysis was used to perform feature extraction [3].

## 1. INTRODUCTION

The approach to large vocabulary recognition we adopt in this paper is a pattern recognition 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. A detailed description of the improved acoustic modeling techniques can be found in [2].

We tested the improved acoustic modeling techniques on speaker-independent recognition of the DARPA Naval Resource Management (RM) 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 95.4%.

## 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 [1]. 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.

### 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. The state observation density of each state of every PLU is represented by a multivariate Gaussian mixture density with a diagonal covariance matrix. Training of the set of sub-word units is accomplished by a modified version of the segmental *k*-means training procedure [1-2].

### 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. In practice, given the set of 47 PLU's, only a small fraction of them appear in words for a given task. However the number of CD PLU's for a set of training data 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. Perhaps the simplest way is to use a *unit reduction rule* [1-2] based on the number of tokens of a particular unit appearing in training.

† Now with CSELT, Torino, Italy.



CH2977-7/91/0000-0161 $1.00 © 1991 IEEE

### 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 3990 read sentences from 109 talkers (30-40 utterances/talker). For consistency in comparing performance results, we used the so-called FEB89 test data which consisted of 300 sentences from 10 new talkers (30 utterances/talker) as distributed by DARPA in February 1989.

To provide a consistent base 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, the threshold resulted in a set of 1090 PLU's. With this set we obtained a word accuracy of 91.3% and a sentence accuracy of 58.7% on the FEB89 test set.

## 3. INTER-WORD UNIT MODELING

Several schemes have been proposed for modeling inter-word, context-dependent units. The methodology we adopted for inter-word unit modeling is described in detail in [4]. A typical implementation discussed in the literature use only double-context and context-independent phones; however we found single-context phones quite helpful in our implementation. Based on the same unit reduction rules [1] 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. Training is based on a modified version of the segmental $k$-means training procedure [1].

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* (FSN) representing the task grammar is converted into an efficient *compiled network* to minimize computation and memory overhead. The reader is referred to [5] for a detailed description of the recognizer implementation. 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.

## 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. We now focus our discussion on an improved front-end feature analysis. Since we are using a *continuous density HMM* (CDHMM) approach for characterizing each of the sub-word units, it is fairly straightforward to incorporate new features. Specifically, we augment the original 24-element feature vector with the second order cepstral derivatives (*delta-delta cepstrum*), the log energy derivative (*delta energy*), and the second order log energy derivative (*delta-delta energy*), giving a 38-dimension feature vector.

We found that the improved feature analysis reduced the word error rate of the FEB89 test result by 30% using the WP grammar [2]. The same modeling techniques was applied directly to the problem of connected digit recognition. Using one model per digit, with 10 states per model and 64 mixture components per state on the TI digit test set, we achieved a string error rate of 1.4% for unknown length strings (with word error rate of 0.4%) and a string error rate of 0.7% for known length strings (with word error rate of 0.2%) [3]. This represents a 50% string error reduction over our best results reported previously.

### Second Order Cepstral Time Derivatives

There are several ways to incorporate the second order time-derivative of the cepstral coefficients. Most of the existing approaches evaluate the delta-delta cepstrum either as the time derivatives of the delta cepstrum or as a regression fit from the original cepstrum. The degree of success in using delta-delta cepstrum was rather mixed. One of the first continuous speech recognition system which used the delta-delta cepstral features was reported by Ney [6]. Ney evaluated the system using speaker independent recognition of the DARPA RM task, and showed a very significant improvement when testing the recognizer without using any grammar. However, for the word-pair grammar, there was no improvement in performance.

In our evaluation, we used a 3-frame window on the delta cepstral coefficients (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-dimension feature vector with 12 additional delta-delta cesptral features giving a 36-dimension feature vector. We then tested this new feature analysis procedure on the resource management task using the WP grammar. Adding the delta-delta cepstral parameters (using both inter-word and intra-word units on the set of 1282 PLU's), improved the word accuracy from 93.0% to 93.8% on the FEB89 test set. 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. We also observed a large variability in speaker performance using models obtained from various iterations. One possible explanation for the difference is that the second order cepstral analysis produces very noisy observations. Another concern is the effectiveness of each of the additional features. In Ney [8], a pre-selected 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.

### Log Energy Time Derivatives

Delta energy has been shown useful in a number of recognition systems. Most systems use both energy and delta energy as features. However, we 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.

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 the 12 delta-delta cepstral coefficients, the delta energy and delta-delta energy parameters to the feature set, we have a 38-element feature vector for every frame $l$.

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. The test results are summarized in Table 1. When compared with the results from the previous section (listed in column 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. When delta-delta energy is added to form a 38-dimension feature vector, we observe that delta-delta energy is not always beneficial. There is no overall improvement (column 3); however all the test speakers achieve over 92% word accuracy. We also show in column 4, the best achievable performance (over 96%) using various feature combinations extracted manually.

| DDCEP | +DENG | +DDENG | BEST |
|---|---|---|---|
| 93.8 | 94.9 | 94.9 | 96.1 |

**Table 1.** Improved feature analysis test results

## 5. POSITION-DEPENDENT UNIT MODELING

For all our experiments so far, we have selected the set of basic speech units based on context dependency. 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. 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). The average unit separation distance is about 9. This indicates that even for units appearing in the same context but in a different environment (i.e. intra-word versus inter-word), the spectral characteristics behave differently. For the 1282 PLU set, the histogram plot is skewed to the left (i.e. less unit separation).

Using the WP grammar, the result based on the 1769 PLU set gave a word accuracy of 95.4% compared with 94.9% for the 1282 PLU set. 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. Similar performance and error patterns were observed for the OCT89 and JUN90 test sets [2]. This type of errors can easily be corrected with a simple set of syntactic and semantic rules.

## 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.

Based on the CDHMM framework, we have shown that careful selection of context-dependent sub-word units in conjunction with detailed acoustic modeling of the set of sub-word units provided high word accuracies. The results reported in the previous sections are summaried in Table 2 for the FEB89 test set using the word-pair grammar. It is noted that the overall error rate reduction is over 60% going from using only context-independent units to combining all the modeling techniques discussed above.

| Test Conditions | Error Rate |
|---|---|
| Context Independent PLU | 14.0% |
| + Context Dependent PLU | 8.7% |
| + Inter-Word CD PLU | 7.0% |
| + Delta-Delta Cesptrum | 6.2% |
| + Energy Time Derivatives | 5.1% |
| + Position Dependent PLU | 4.6% |

**Table 2.** Summary of word error rates

To date, the criterion for selecting the set of fundamental units is based on the rule of observing enough occurrences of each unit in the training data. With the incorporation of position-dependent units we observed that the performance of the system was relatively insensitive to the unit reduction threshold used in deciding the set of units to model. To illustrate this point we show, in Table 3, the word error rates as a function of the threshold for both the word-pair case and the no grammar case for the FEB89 test set.

| Threshold | 30 | 25 | 20 | 15 | 10 |
|---|---|---|---|---|---|
| No. of Units | 1769 | 2135 | 2534 | 2985 | 3863 |
| WP Error(%) | 4.6 | 4.7 | 4.6 | 4.7 | 5.3 |
| NG Error(%) | 20.3 | 19.8 | 19.4 | 20.6 | 20.9 |

**Table 3.** Word error rates as a function of threshold

It is noted that all the results presented in this paper were obtained using a compiled network beam search algorithm with search guided by the correct sentence [5]. More recently, we found that the actual recognition results without the help of the *guided search* degraded the results by 0.3% to 1.0% [5]. However the relative performance comparison remains the same. In order to make a consistent comparison we still use the guided search recognition results throughout this study.

Even high accuracy has been achieved, there are still some open issues that need to be addressed. 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 unit training; (2) Lexical modeling to deal with lexical variability in baseform pronunciation; (3) Improved feature selection so that only those features useful to discrimination are included in the feature vector; and (4) Smoothing, interpolation, corrective training and adaptation of CDHMM parameters [7].

Quick speaker adaptation was shown effective for isolated word recognition based on Bayesian adaptation of the parameters of a multivariate Gaussian density for CDHMM [8]. For large vocabulary speech recognition, the amount of training data is always not sufficient in characterizing the models for the set of sub-word units. Some form of parameter smoothing and model sharing is required. In a recent study [7], it was found that Bayesian learning serves as a unified approach for parameter smoothing, corrective training, speaker clustering and speaker adaptation. A theory on Bayesian learning of CDHMM with a multivariate Gaussian mixture density has been developed [7]. These parameter smoothing and speaker clustering techniques have been applied to the RM task and achieved a 12% error reduction on the more difficult JUN90 test data. When applied to speaker adaptation on the RM task, using only 2 minutes of speaker-specific training data, a 32% error reduction was attained over our best speaker independent results [7]. Study is also underway on using the same Bayesian learning principle for corrective training.

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. For example, for the RM task, we have developed a language decoder (decoupled from the acoustic decoder) that incorporates a set of simple linguistic rules. Our preliminary results [9] indicate that the sentence *semantic accuracy* for the FEB89 test set improved from 75% to close to 95% (with a 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 a 90% word accuracy). The effectiveness of this approach 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 and J. G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, 4, 1990.

[2] C.-H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg "Improved Acoustic Modeling for Continuous Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, Somerset, PA., June 1990.

[3] J. G. Wilpon, C.-H. Lee and L. R. Rabiner, "Improvements in Connected Digit Recognition Using Higher Order Spectral Features," *Proc. ICASSP 91*, Toronto, Canada.

[4] 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.

[5] R. Pieraccini, C.-H. Lee, E. Giachin and L. R. Rabiner, "Complexity Reduction in a Large Vocabulary Speech Recognizer," *Proc. ICASSP 91*, Toronto, Canada, May 1991.

[6] 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.

[7] J.-L. Gauvain and C.-H. Lee, "Bayesian Learning of Gaussian Mixture Density for Hidden Markov Models," submitted for publication.

[8] C.-H. Lee, C.-H. Lin and B.-H. Juang, "A Study on Speaker Adaptation of the Parameters for Continuous Density Hidden Markov Models", in *Proc. ICASSP90*, to appear in *IEEE Trans. on Acoustic, Speech and Signal Proc.*, 1991.

[9] R. Pieraccini and C.-H. Lee, "Factorizing Language Constraints in Speech Recognition," submitted for publication.