ACOUSTIC MODELING OF SUBWORD UNITS FOR LARGE VOCABULARY SPEAKER INDEPENDENT SPEECH RECOGNITION

Chin-Hui Lee, Lawrence R. Rabiner, Roberto Pieraccini† *and Jay G. Wilpon*

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

ABSTRACT

The field of large vocabulary, continuous speech recognition has advanced to the point where there are several systems capable of attaining between 90 and 95% word accuracy for speaker independent recognition of a 1000 word vocabulary, spoken fluently for a task with a perplexity (average word branching factor) of about 60. There are several factors which account for the high performance achieved by these systems, including the use of hidden Markov models (HMM) for acoustic modeling, the use of context dependent sub-word units, the representation of between-word phonemic variation, and the use of corrective training techniques to emphasize differences between acoustically similar words in the vocabulary. In this paper we describe one of the large vocabulary speech recognition systems which is being developed at AT&T Bell Laboratories, and discuss the methods used to provide high word recognition accuracy. In particular, we focus on the techniques used to obtain acoustic models of the sub-word units (both context independent and context dependent units), and discuss the resulting system performance as a function of the type of acoustic modeling used.

INTRODUCTION

In the past few years there have been proposed a number of systems for large vocabulary speech recognition which have achieved high word recognition accuracy [1-6]. Although a couple of the systems have concentrated on either isolated word input [6], or have been trained to individual speakers [5,6], most current large vocabulary recognition systems have the goal of performing speech recognition on fluent input (continuous speech) by any talker (speaker independent systems).

The approach to large vocabulary speech recognition we adopt in this study is a pattern recognition based approach. For a detailed description of the system we have developed, the reader is referred to [7]. The basic speech units in the system 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 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.*

The focus of this paper is a discussion of various methods used to create a set of acoustic models for characterizing the PLU's used in large vocabulary recognition (LVR). The set of context independent (CI) units we used in this study is a fixed set of 47 phone-like units (PLU's), in which each PLU is associated with a linguistically defined phoneme symbol. We model each CI PLU using a continuous density hidden Markov model (CDHMM) with a Gaussian mixture state observation density. Each word model is defined as the concatenation of the PLU models according to a fixed lexicon defined by the set of 47 associated phoneme symbols. We also consider a set of context dependent (CD) units which includes PLUs' defined by left, right and both left and right context.

† On leave from CSELT, Torino, Italy.

We tested the recognition system on the DARPA Naval Resource Management task using the word-pair (WP) grammar in a speaker independent mode. In the case of context independent acoustic modeling, we varied the maximum number of mixtures in each state from 1 to 256 and found that the word accuracy increased from 61% to 90% which indicates that sufficient acoustic resolution is essential for improved performance. The 90% word accuracy is the highest performance reported based on context independent units. When intraword context dependency modeling is incorporated, we improved out performance to 93% word accuracy.

ACOUSTIC MODELING OF SUB-WORD UNITS

A block diagram of the LVR system we are using is shown in Figure 1. The speech input is first filtered from 100 Hz to 3.8 kHz, and sampled and digitized at an 8 kHz rate. The digitized speech is then pre-emphasized. A 10^{th} LPC analysis is performed on a Hamming-windowed speech segment of 30 msec, and every 10 msec a feature vector consisting of 12 liftered cepstral coefficients and 12 corresponding time derivatives is generated. Temporal features such as log energy and various durational features can also be used as part of the observation vector for training and recognition.

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

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.

Based on the choice of PLU's as the set of sub-word units for recognition, the only missing knowledge in the system in order to perform recognition is the set of sub-word PLU models. We now describe the techniques used to obtain acoustic models of the sub-word units (both context independent and context dependent).

SUB-WORD HIDDEN MARKOV MODELS

The units chosen in our research were a set of 47 PLU's corresponding to a set of 47 English Phoneme symbols. The speech units are modeled left-to-right, *continuous density* hidden Markov models. In our implementation, we set the number of states in each model at a fixed value of 3. This, of course, implies that the shortest tokens of the sub-word unit last at least 3 frames. Within each state of the HMM, the random spectral vector is represented by a Gaussian mixture density. Each mixture component has a spectral mean and variance which is highly dependent on the spectral characteristics of the sub-word unit (i.e. highly localized in the total acoustic space).

TRAINING OF PLU MODELS

In order to train a set of sub-word PLU's for LVR, i.e. to estimate the "optimal" parameters of the PLU models, we need a labeled training set of continuous speech, where the labeling consists of an ASCII representation of the spoken text within each utterance. To train the PLU models we represent each sentence in the training set as a (not necessarily unique) sequence of sub-word units with the option of silence between any pair of words, and at the beginning and/or end of each sentence. Hence if we have the sentence, S, which consists of the words $W_{S1}\, W_{S2} \ldots W_{SJ}$, then we can represent the sentence in terms of PLU's by first modeling the sentence as a series of optional silences followed by the specified words. Finally each lexical entry is replaced by its sequence of sub-word PLU's, as expressed in the lexicon. Finally we allow model multiples for each PLU so we replace each canonic PLU by 1 or more models in parallel. The network created by embedding the multiple phone models into each lexical entry, and by embedding the multiple lexical entries into each word, and finally by embedding the word models into each sentence is then used to match the spectral representation of the input via a Viterbi matching procedure. By backtracking we can determine which phone model (in the case of multiple phone models) and which lexical entry (in the case of multiple lexical entries) gave the best match and use these as the best representation of the input utterance.

By using the above procedure on all utterances within a given training set, we can estimate the PLU model parameters via a variant on the segmental k-means training procedure [8]:

1. **Initialization** — linearly segment all training utterances into units and HMM states; assume a single lexical entry per word (any one can be used) and a single model per sub-word unit.

2. **Clustering** — all frames (observation vectors) corresponding to a state S_j in all occurrences of a given sub-word unit are partitioned into M_j clusters (using standard VQ design methods).

3. **Estimation** — the mean vectors, μ_{jm}, the (diagonal) covariance matrices, $\mathbf{U}_{jm}$, and the mixture weights, c_{jm}, are estimated for each cluster m ($1 \leq m \leq M_j$) in state S_j.

   (By cycling steps 2 and 3 through all sub-word units and through all states of each sub-word unit, a set of HMM's is created.)

4. **Segmentation** — the PLU set of HMM's is used to (re)segment each training utterance into units and HMM states via Viterbi decoding; multiple lexical entries per word as well as multiple models per PLU are now allowed.

5. **Iteration** — steps 2-4 are iterated until convergence, i.e. until the average likelihood of the matches essentially stops increasing.

By applying the segmental k-means training procedure to a set of 4360 sentences from 109 different talkers, we obtain a total of 155000 PLU labels. The segments labeled silence (*h#*) have the most occurrences (10638 or 6.86% of the total) and *nx* (syllabic *n*) have the fewest occurrences (57 or 0.04% of the total). In terms of average likelihood scores, silence (*h#*) had the highest score (18.5) followed by *f* (17.7) and *s* (15.4), while *ax* had the lowest score (7.1) followed by *n* (8.3) and *r* (8.4). It is interesting to note that the PLU's with the 3 lowest average likelihood scores (*ax*, *n* and *r*) were among the most frequently occurring sounds (*r* was second, *n* sixth and *ax* fourth in frequency of occurrence). Similarly some of the sounds with the highest likelihood scores were among the least occurring sounds (e.g. *oy* was fourth according to likelihood score but 21st according to frequency of occurrence). These results almost obey a type of Zipf's law which, in terms of the PLU statistics, states that there is an inverse relationship between frequency of occurrence and ability to model the sound.

CREATION OF CONTEXT INDEPENDENT PLU MODELS

The use of CI PLU's has several advantages, namely: (1) the PLU models are easily trained, (2) no smoothing is required, (3) the units themselves are relatively insensitive to the context from which the training tokens are extracted, and (4) the units are readily generalized to new contexts, e.g. new vocabulary sets, new word pronunciations etc. However, the use of CI PLU's also leads to two serious problems, namely: (1) the CI PLU's do not represent the unit well in all contexts, and (2) the CI PLU's do not provide high recognition performance for large vocabulary recognition tasks, i.e. no one has achieved over 90% word recognition accuracy for vocabularies of 1000 or more words based solely on using context independent PLU's.

There are at least three reasonable solutions to the above problems, namely: (1) improve the acoustic resolution of the context independent PLU models by either modifying the model structure or by using more mixture components in each state, (2) increase the number of models for each context independent PLU thereby reducing the acoustic variability within each model. and (3) create a set of context dependent PLU models and modify the word lexicon to account for the new set of units.

Perhaps the simplest way of improving the acoustic resolution of the context independent PLU models is to use more detailed representations of each unit. In this paper, we focus our discussion on the technique of increasing the number of mixture densities per state. The ultimate limitation here is the amount of training data per unit. Although some units have a large number of occurrences in the training set, the less frequently occurring units will not have enough occurrences to justify a large number of mixtures per state. The obvious solution here is to use a strategy in which the number of mixtures per state is a function of the size of the training set and to stop increasing the number of mixtures for a given unit when it exceeds some critical value. We will show later that increasing acoustic resolution in sub-word modeling effectively improves recognition performance.

CREATION OF CONTEXT DEPENDENT PLU MODELS

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]. The more general case of both left and right context dependent PLU's represents each phone $p \rightarrow p_L - p - p_R$ where p_L is the preceding phone (possibly silence) and p_R is the following phone (possibly silence). For the time being, we assume that we do not cross word boundaries when creating CD models.

The way in which we create CD PLU models is as follows: we first convert the lexicon from CI units to CD units, we then train the set of CD PLU's using the same procedure as used for the CI PLU's, i.e. use the segmental *k*-means training on the expanded set of PLU's until convergence.

The above training procedure leads to one major problem, namely that the number of occurrences of some of the CD units is insufficient to generate a statistically reliable model. There are several ways of dealing with this problem. Perhaps the simplest way is to use a unit reduction rule of the form: if $c(p_L - p - p_R) < T$, then

1. $p_L - p - p_R \longrightarrow \$ - p - p_R$ if $c(\$ - p - p_R) > T$
2. $p_L - p - p_R \longrightarrow p_L - p - \$$, if $c(p_L - p - \$) > T$
3. $p_L - p - p_R \longrightarrow \$ - p - \$$

where $c(p_1 - p_2 - p_3)$ is the count in the training set associated with the ordered triplet (p_1, p_2, p_3) ($ is a don't care or wild card phone), and T is the count threshold for applying the reduction rule sequentially through the 3 cases.

To illustrate the sensitivity of the CD PLU set to the threshold of occurrences, T, Table 1 shows the counts of left and right context PLU's, left context PLU's, right context PLU's, and context independent PLU's for the 109 talker DARPA training set of 4360 sentences, as a function of T. It can be seen that for a threshold of 50, which is generally adequate for estimating the HMM parameters, there are only 365 intraword left and right context PLU's (out of a possible 103,823 combinations), and even for a threshold of 1, there are only 1778 intraword left and right context PLU's; hence only a very small percentage of the possible left and right context PLU's occur in this 4360 sentence set.

Count Threshold	Number of Left and Right Context PLU's	Number of Left Context PLU's	Number of Right Context PLU's	Number of Context Independent PLU's	Total Number of CD PLU's
50	378	158	171	47	754
40	461	172	188	47	868
30	639	199	205	47	1090
20	952	212	234	46	1444
10	1302	243	258	44	1847
5	1608	265	270	32	2175
1	1778	279	280	3	2340

Table 1. Counts of Intraword CD Units as a Function of Count Threshold (T)

A second way of handling the insufficiency of the data for creating statistically reliable CD PLU's is to smooth the CD models with CI models via a technique like deleted interpolation [10]. In order to use deleted interpolation both the CD and the CI models need to be created based on a common codebook (e.g. discrete observation probabilities) or based on a common set of Gaussian densities (e.g. the mixed density method). If this is the case then if we denote the spectral density for the CI unit $\$-p-\$$ in state j as B_j^{CI}, and the spectral density for the CD unit p_L-p-p_R in state j as B_j^{CD}, then we create the smoothed spectral density $\bar{B}_j^{\mathrm{CD}}$ as a linear combination of B_j^{CD} and B_j^{CI}. The weight is estimated directly from training data which is deleted (withheld) from the training data used to create B_j^{CD} and B_j^{CI}. The forward-backward algorithm can be used directly to estimate the weight λ [10].

The key to the success of the deleted interpolation procedure is the commonality of the spectral densities used for the CD and CI units. A slightly different way of exploiting this type of smoothing is to use the mixed density method but localized to each CI PLU. Thus for designing each CD PLU, we assume that within each state the means and covariances of each mixture are the same as those used for the CI PLU model; however we adjust the mixture gains based on the actual occurrences of each CD PLU in the training set. We can also then apply a form of interpolation which is similar to that of deleted interpolation to the mixture gains by smoothing them with the CI mixture gains, i.e.

$$\bar{c}_j(m)^{p_L-p-p_R} = \lambda c_j(m)^{p_L-p-p_R} + (1-\lambda)c_j(m)^{\$-p-\$} \qquad (1)$$

where λ is again estimated from counts of training tokens where the CD model provides a better fit than the CI model. This type of smoothing is especially effective for models created from a small number of training tokens (e.g. less than 30).

We have therefore considered two types of modeling algorithms for creating CD PLU's, based on the above discussion. The first procedure, which we refer to as CD1, sets a threshold on the minimum number of CD PLU occurrences in the training set and then, independent of the CI phone set, builds a new set of CD models. The second procedure, which we refer to as CD2, uses the modified training/smoothing procedure to tie the mixture means and covariances, and allows the use of a simple interpolation scheme (Eq. (1)). We will present results of both these CD PLU model creation procedures in the next section.

EXPERIMENTS AND RESULTS

As described above, we use a finite state network (FSN) to represent the language model of the recognition task. In implementing the FSN, we can allow deterministic $(0 \mid 1)$ or probabilistic (bigram probabilities in context) connections between words, and can even incorporate word insertion penalties. The FSN for the DARPA naval resource management task is given in Figure 2. The vocabulary consists of 991 words which have been sorted into 4 non-overlapping groups, namely

$\{BE\}$ = set of words which can begin a sentence or end a sentence, $|BE| = 117$

$\{B\bar{E}\}$ = set of words which can begin a sentence but which cannot end a sentence, $|B\bar{E}| = 64$

$\{\bar{B}E\}$ = set of words which cannot begin a sentence but can end a sentence, $|\bar{B}E| = 448$

$\{\bar{B}\,\bar{E}\}$ = set of words which cannot begin or end a sentence, $|\bar{B}\,\bar{E}| = 322$.

To account for interword silence (again optional) we expand each word arc bundle (e.g. node 1 to node 4) to individual words followed by optional silence, as shown at the bottom of Fig. 2. Depending on the preceding decoded word, word bigram probabilities are trivially inserted at the beginning of every word arc, and word insertion penalties are similarly easily used at the word output nodes (5, 6, 7 & 8).

For all the experiments to be reported on in the following we used the FSN of Fig. 2 with either specified allowable word pair combinations (WP, word pair grammar), or with any transition between all pairs of words being equally likely (NG, no grammar case). In our tests, as prescribed by DARPA standard reporting procedures [11], we have used mainly the WP grammar; however we present results on the NG case for comparison with results of other researchers.

EXPERIMENTAL SETUP

For most of our tests we used the training material provided by DARPA. The speech database was provided by DARPA at a 16 kHz sampling rate. We filtered and down-sampled the speech to an 8 kHz rate before analysis. The first training set, which we call TR1, consists of a set of 3200 sentences from 80 talkers (40 sentences/talker). We used three separate testing sets to evaluate the recognition system trained from TR1 (80), including:

1. 150 sentences from 15 talkers (10 sentences/talker) not included in the 80 talker training set. This set is identical to the one used by Lee at CMU to initially evaluate the SPHINX system [1], and we call this set TS1 (150).
2. 300 sentences from 10 other talkers (30 sentences/talker) as distributed by DARPA in February 1989. We call this set TS2 (FEB 89).
3. A set of 160 randomly selected sentences from the set of 3200 training sentences (2 randomly selected sentences from each of the 80 training talkers) which we created to check on the closed set performance of the system. We call this set TS3 (TRAIN).

A second training set was also used consisting of 4360 sentences from 109 talkers (40 sentences per talker). We call this training set TR2 (109). The 109 talker set overlapped the 80 talker set (TR1) in that 72 talkers were common to both sets. The remaining 37 talkers in TR2 partially overlapped the talkers in TS1 (150). Hence the only independent test set for TR2 was TS2 (FEB 89).

BEAM SEARCH RECOGNITION ALGORITHM

The way in which the recognizer was implemented was to use the FSN of Fig. 2 directly and to keep track of the accumulated likelihood score to each node in the network. That is we expand each word bundle into individual words, expand each word into one or more sequences of PLU's (via the lexicon), and expand each PLU into HMM states of the corresponding model (or models). Thus the network of Fig. 2 has on the order of 20,000 HMM states and word junction nodes to keep track of at each frame of the input. To reduce computation, a frame-synchronous beam search algorithm [12] is used in which the best accumulated likelihood, L^*, is determined, at each frame, and based on a threshold, Δ, all nodes whose accumulated likelihoods are less than $(L^* - \Delta)$ are eliminated from a list of active nodes (i.e. paths from these nodes are no longer followed). A key issue is then how to set Δ so as to eliminate a high percentage of the possible paths, but not to eliminate the ultimate best path. The problem with a fixed value of Δ is that in regions where the word matches are not very good (e.g. function words) you need a relatively large value of Δ (because of ambiguities which won't be resolved until some content words are included) but in regions where the word matches are excellent (e.g. content words, names of ships, etc.) you can use a fairly small value of Δ and still not eliminate the best path.

The time for computation varied almost linearly with Δ; hence the penalty paid for a large Δ is storage and computation time, but the reward is that the best string is obtained. Clearly these results show the need for an adaptive beam width algorithm which can reduce its size during regions of good word matches, and increase its size during regions of relatively poor word matches. Such a procedure does not yet exist.

RESULTS WITH CI PLU MODELS

For the basic CI 47 PLU set we used training set TR1 and iterated the segmental k-means procedure until convergence (10 iterations from a uniform initialization). We then used the resulting segmentation into units to design model sets with the nominal maximum number of (diagonal covariance) mixtures per state varying from 1 to 256 in several steps. The resulting models were run on the 3 test sets for the 991 word DARPA task using the WP grammar, and the word recognition accuracies as a function of the nominal maximum number of mixtures per state are listed in Table 2. It can be seen that large improvements in word recognition accuracy are obtained as the number of mixtures/state, M, is increased from 1 to 18 (about 20% for each of the 3 test sets). However as M is increased even further, from 18 to 75, word accuracies increase much less rapidly (by 2.2% for TS1 for 128 mixtures/state, 4.6% for TS2 and 6.9% for TS3) for all 3 test sets. Beyond $M = 75$, performance essentially bottoms off for both independent test sets (TS1 and TS2) and increases by 2.0% for TS3 (the training set). This result shows that by increasing acoustic resolution, performance continues to increase so long as there is sufficient training data (as is the case for 47 CI PLU's). It is also noted that for TS1 (open test), we achieve close to 90% word accuracy by simply using the set of CI PLU's.

Number of Mixtures Per State	RECOGNITION TEST SET		
	TS1 (150)	TS2 (FEB 89)	TS3 (TRAIN)
1	64.7	61.3	67.8
3	76.7	72.4	79.2
6	82.9	78.1	82.9
9	83.8	79.6	85.6
18	87.5	80.8	88.5
36	88.3	83.9	90.1
75	89.7	85.4	93.3
128	89.9	85.0	94.2
256	89.6	86.0	95.3

Table 2. Word Recognition Accuracies (%) for TS1, TS2, TS3
Using the 47 CI PLU Models Derived from the 80 Talker Training Set

RESULTS WITH CD PLU MODELS

Using the CD1 method of creating CD PLU's (i.e. by setting a threshold of 50 occurrences of each intraword left and right context dependent PLU and backing down to intraword left and/or right context dependent PLU's, and/or context independent PLU's), a set of 638 CD PLU's was created from the 80 talker training set, TR1. The composition of the 638 CD PLU set was: 304 left and right context PLU's, 150 right-context PLU's, 137 left-context PLU's, and all 47 context independent PLU's.

For this 638 CD PLU set, models were created with 9, 16, and 32 mixtures/states. Initial model estimates were obtained from the 47 CI PLU segmentations, and the segmentation was then iterated 2-4 times for each different size model. Recognition results on the three test sets are given in Table 3. It can be seen that the word recognition accuracies increase by 4.2% for TS1, 4.7% for TS2 and 5.4% for TS3 as the number of mixtures/state goes from 9 to 32 (32 was the largest size model that was reasonable to try on this data).

Nominal Number of Mixtures per State	TEST SET		
	TS1 (150)	TS2 (FEB 89)	TS3 (TRAIN)
9	88.5	85.2	93.3
16	92.3	89.7	97.9
32	92.7	89.9	98.7

Table 3. Word Recognition Accuracies (%) for 638 CD1 PLU Set

Next we created context dependent PLU sets using the CD2 method where we used the 256 mixture/state CI PLU model as the base model and varied only the mixture gains in each state of each CD PLU. CD PLU sets were created with count thresholds of infinity (47 CI PLU set), 50 (638 CD PLU set), 30 (915 CD PLU set), 10 (1759 CD PLU set) and 1 (2340 CD PLU set) using the 80 talker training set. The resulting models were tested based on raw mixture gains, as estimated entirely from training set tokens of each CD PLU, and with smoothed mixture gains, as estimated by interpolation of the CI PLU mixture gains with the CD PLU mixture gains (Eq.(1)). Estimates of the smoothing factor, λ, for each state of each CD PLU were obtained *entirely* from training set data. The results on these sets of units are given in Table 4, both for the word pair (WP) grammar (Table 4a), and the no grammar

(NG) case (Table 4b).

The results in Table 4a, for the WP grammar, show that for count thresholds of 1 and 10, the results obtained from smoothed parameters are better than those from the raw parameters for both TS1 and TS2 data. This is to be expected since the amount of training data for many of the CD PLU's (i.e. those with less than 10 occurrences) is inadequate to give good mixture gain estimates, and the smoothing helps a good deal here. For count thresholds of 30 and 50 there is a small performance advantage for the raw parameters models (i.e. 1.3% for TS1 for count of 30, 0.6% for TS1 for count of 50, 0.3% for TS2 for count of 30, −0.1% for TS2 for count of 50), but here the differences in word accuracy are relatively small.

The best performance, on the WP grammar, for the CD2 method of creating CD PLU's is 93.3% for TS1 (both 2340 and 1759 smoothed parameters CD PLU sets) and 90.9% for TS2 (638 smoothed parameter CD PLU set). These results represent a 0.6% improvement for TS1 and a 1.0% improvement over the 638 CD PLU set created with 32 mixtures/state from the CD1 method (as shown in Table 3). Although the level of improvement is relatively small, there is a consistent trend to obtaining slightly higher performance with the CD2 method of creating CD PLU's.

The results in Table 4b, for the NG case, again show improved performance for the smoothed parameters case (over the raw parameters model) for both count thresholds of 1 and 10 for TS1 and TS2 data. For count thresholds of 30 and 50, we again see that the smoothing tends to slightly degrade word recognition accuracy. The best performance, on the NG grammar, for the CD2 method is 72.1% for TS1 and 68.8% for TS2 for the case of 2340 CD PLU's with smoothed parameter estimates.

Count Threshold	Number of CD PLU's	Raw Parameters Test Set			Smoothed Parameters Test Set		
		TS1	TS2	TS3	TS1	TS2	TS3
1	2340	91.4	88.2	97.6	93.3	89.9	97.4
10	1759	92.6	89.3	97.4	93.3	90.6	97.2
30	915	93.2	90.3	97.1	91.9	90.0	97.0
50	638	92.9	90.8	97.0	92.3	90.9	97.0
	47	89.6	86.0	95.3	–	–	–

(a) Word Accuracies (%) Based on the Word-Pair Grammar

Count Threshold	Number of CD PLU's	Raw Parameters Test Set			Smoothed Parameters Test Set		
		TS1	TS2	TS3	TS1	TS2	TS3
1	2340	67.8	65.6	91.2	72.1	68.8	90.1
10	1759	69.6	66.7	91.0	69.8	68.6	89.6
30	915	68.6	67.9	88.7	67.1	66.2	87.9
50	638	67.1	66.9	89.1	67.4	66.2	88.6
	47	60.2	60.0	82.6	–	–	–

(b) Word Accuracies (%) Based on the NG Grammar

Table 4. Recognition results Based on CD2 PLU's Derived from the 80 Talker Training Set

SUMMARY OF RESULTS

A summary of the best performances of the three types of PLU units, CI PLU's, CD1 PLU's and CD2 PLU's, discussed in this paper is given in Table 5 which shows, for each test set, the sentence accuracy, the word correct, word substitution, word deletion, word insertion, and word accuracy rates. The results are given for the WP grammar based on the 80 talker training set (TR1).

The results show a steady improvement in performance in going from 47 CI PLU's to 638 CD PLU's for all 3 test sets. Although the CD2 method of creating CD PLU's provides small improvements in performance (in terms of word accuracy) for TS1 and TS2 10.6% and 1.0%), the sentence accuracies are not higher with this method. (In fact sentence accuracy is 4.7% higher for the CD1 method, for TS1, than for the CD2 method; for TS2 the sentence accuracies are comparable; for TS3, the training set, sentence accuracy is 7.4% higher for the CD1 method).

Number of PLU's	Context	Test Set	Sentence Accuracy (%)	Word Accuracies and Error Rates (%)				
				Correct	Substitution	Deletion	Insertion	Accuracy
47	CI	TS1	52.4	91.0	5.9	3.1	1.1	89.9
47	CI	TS2	45.0	87.0	4.4	4.4	1.0	86.0
47	CI	TS3	69.4	95.6	1.7	2.7	0.3	95.3
638	CD1	TS1	70.7	94.8	4.1	1.1	2.0	92.7
638	CD1	TS2	56.3	90.9	6.5	2.6	1.0	89.9
638	CD1	TS3	88.7	98.8	0.1	1.1	0.1	98.7
1759	CD2	TS1	66.0	94.0	3.8	2.3	0.7	93.3
638	CD2	TS2	56.7	91.7	5.3	3.0	0.8	90.9
2340	CD2	TS3	81.3	97.7	0.6	1.7	0.1	97.6

Table 5. Detailed Performance Summary for WP Grammar for CI and CD Unit Sets, Based on 80 Talker Training Set

DISCUSSION

The results presented in the previous section show that proper acoustic modeling of the basic sub-word recognition units is essential for high recognition performance. Although the performance of the resulting system on the DARPA Resource Management System is good, there is still a great deal that needs to be done to make such a recognition system practically useful. In this section we first discuss how the results presented in this paper compare to those of other researchers working on the same task. Then we discuss the areas that we feel would be most fruitful for further research.

COMPARISON OF RESULTS

Since a large number of research groups are using the DARPA Resource Management Task as a standard training/test set, it is relatively straightforward to make direct comparisons of performance scores. However, before doing so, it is appropriate to point out that, aside from system differences, there are often a number of methodology differences that could significantly affect the results. When appropriate we will point out these differences.

For TS1 (150) the most appropriate comparison is the results of Lee and his colleagues at CMU, since Lee essentially defined the data that went into TS1 [1]. The SPHINX System, which uses a multiple VQ front end (i.e. a discrete observation density rather than the continuous mixture density used here), has been in development for about 5 years, and has learned how to exploit durational information (words) as well as function word dependent phones. The SPHINX system also uses a somewhat larger training set (105 talkers, 4200 sentences) than used here.

Based on the results presented in [1], using 3 codebooks, duration, function word phones, and generalized triphones (similar to CD PLU's discussed here), Lee obtained 93.7% word accuracy with the WP grammar on TS1 (150), and 70.6% word accuracy with the NG grammar [1]. These results are comparable to the 93.3% word accuracy obtained for a 1759 CD PLU set on TS1 with the WP grammar and 72.1% word accuracy obtained for a 2340 CD PLU set with the NG grammar, as shown in Table 4.

More recently, Lee *et al.* [13] have incorporated between-word training of the context dependent units (as well as between-word decoding) and a form of corrective training (a word discrimination procedure) to significantly improve recognition performance. Their current results are 96.2% word accuracy for TS1 with the WP grammar and 81.9% with the NG grammar using all the above techniques. This performance represents the highest-to-date reported word accuracy on any fluent speech, speaker independent, large vocabulary task.

For comparisons of performance on the TS2 (FEB 89) test set, performance scores from CMU (Lee *et al.*), SRI (Murveit *et al.*), LL (Paul) and MIT (Zu *et al.*) were recently reported on at a DARPA *Speech and Natural Language Workshop* (February, 1989). The reported word and sentence accuracies along with our results are listed in the following:

Lab	Training Set Size	Word Accuracy	Sentence Accuracy
CMU	109 Talkers	93.9	65.7
AT&T	109 Talkers	91.6	57.7
SRI	109 Talkers	91.2	57.3
LL	109 Talkers	90.2	55.7
MIT	72 Talkers	86.4	45.3

It should be noted that the results reported by CMU, SRI and LL all used both intra-word and inter-word context dependent units whereas those reported by AT&T (as presented here) and MIT did not use inter-word units. Further the MIT system only used a set of 75 CI units including 32 stressed and 32 unstressed vowels, which accounts for the somewhat lower performance scores than the other systems. The results show that the CMU system outperforms the SRI, AT&T and LL systems by about 2.5% for the WP grammar in word accuracy. This result is primarily due to the use of corrective training and inter-word units.

OVERALL ERROR PATTERNS

A detailed analysis of the types of word errors made for the best case of each of the 3 test sets shows the following:

TS1 – 48 substitution errors, 37 involving a function word; 29 deletion errors (*the*(15), *a*(4), *is*(3), *in*(2)) with all 29 errors involving function words; 9 insertion errors (*the*(2)) with 4 of them being function words.

TS2 – 136 substitution errors (*what* → *was* (7)) with 91 involving a function word; 76 deletion errors (*the*(37), *is*(8), *in*(7)) with 70 involving a function word; 20 insertion errors (*of*(3), *is*(3)) with 13 involving function words.

TS3 – 2 substitution errors with 1 involving a function word; 15 deletion errors (*the*(9), *a*(3)) with 14 involving a function word; 2 insertion errors with 1 being a function word.

The message here is clear. We need to significantly improve modeling of function words which involve on the order of 60-75% of the substitution, insertion and deletion errors that are made. The problems here are numerous in that the function words are extremely context sensitive. Several possibilities will have to be investigated including function word dependent PLU's (as used by Lee [1]), inter-word training of CD PLU's, multiple models of function word PLU's, and finally multiple lexical entries for these words.

AREAS FOR FURTHER RESEARCH

Based on the results presented here, as well as those given in the literature, it is clear that there are many areas that must be studied in order to significantly improve word recognition accuracy. These acoustic and lexical modeling issues include:

1. Improved spectral and temporal feature representation
2. Improved function word and function phrase modeling
3. Incorporation of inter-word CD PLU models into training and recognition
4. Some form of corrective training to improve word discrimination capability
5. Acoustic design of lexicon to improve word and phrase modeling

Each of these areas will be investigated in the near future.

SUMMARY

In this paper we have discussed methods of acoustic modeling of basic speech sub-word units so as to provide high word recognition accuracy. We showed that for a basic set of 47 context independent phone-like units, word accuracies on the order of 86-90% could be obtained on a 1000 word vocabulary, in a speaker independent mode, for a grammar with a perplexity of 60, on independent test sets. When we increased the basic set of units to include context dependent units, we were able to achieve word recognition accuracies of from 91 to 93% on the same test sets. Based on outside results and some of our own preliminary evaluations, it seems clear that we can increase word recognition accuracies by about 2-3% based on known modeling techniques. The challenge for the immediate future is to learn how to increase word recognition accuracies to the 99% range, thereby making such systems useful for simple database management tasks.

REFERENCES

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

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

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

4. 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.

5. R. Schwartz *et al.*, "The BBN BYBLOS Continuous Speech Recognition System," *Proc. Speech and Natural Language Workshop*, pp. 94-99, Feb. 1989.

6. F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," *Proc. IEEE*, Vol. 73, No. 11, pp. 1616-1624, November 1985.

7. C.-H. Lee, L. R. Rabiner, R. Pieraccini and J. G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," *submitted for publication.*

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. F. Jelinek, and R. L. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data," *Pattern Recognition in Practice*, E. S. Gelsema, and L. N. Kanal, Ed., North-Holland Publishing Co., Amsterdam, pp. 381-397, 1980.

11. D. Pallett, "Test Procedures for the March 1987 DARPA Benchmark Tests," *DARPA Speech Recognition Workshop*, pp. 75-78, March 1987.

12. B. Lowere, and D. R. Reddy, "The HARPY Speech Understanding System," *Trends in Speech Recognition*, W. Lee, Ed., Prentice-Hall Inc., pp. 340-346, 1980.

13. K. F. Lee, H. W. Hon, and M. Y. Hwang, "Recent Progress in the SPHINX Speech Recognition System," *Proc. DARPA Speech, and Natural Language Workshop*, pp. 125-130, Feb. 1989.

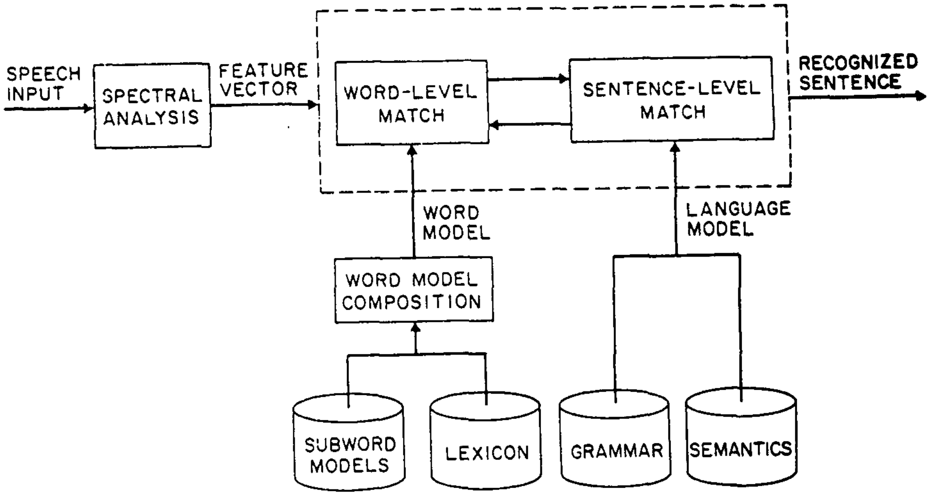

Figure 1. A block diagram of the large vocabulary speech recognition system.

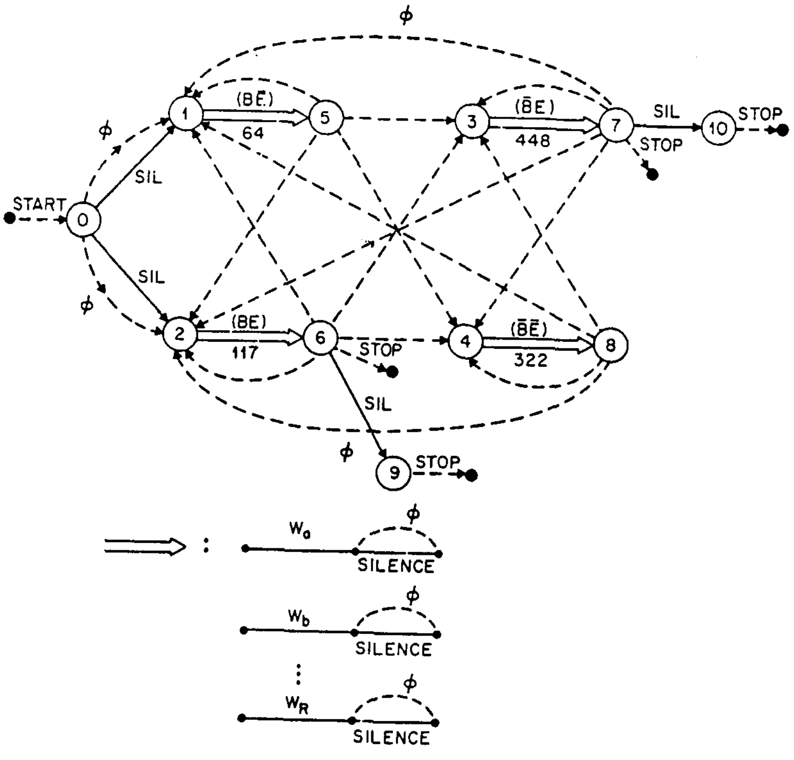

Figure 2. FSN of the DARPA task syntax in which words are partitioned into 4 non-overlapping sets and optional silence is allowed at the beginning and end of the sentence, as well as between pairs of words.