THE MIT SUMMIT SPEECH RECOGNITION SYSTEM: A PROGRESS REPORT*

Victor Zue, James Glass, Michael Phillips, and Stephanie Seneff

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

Recently, we initiated a project to develop a phonetically-based spoken language understanding system called SUMMIT. In contrast to many of the past efforts that make use of heuristic rules whose development requires intense knowledge engineering, our approach attempts to express the speech knowledge within a formal framework using well-defined mathematical tools. In our system, features and decision strategies are discovered and trained automatically, using a large body of speech data. This paper describes the system, and documents its current performance.

INTRODUCTION

For slightly over a year, we have focused our research effort on the development of a phonetically-based spoken language understanding system called SUMMIT. Our approach is based on the belief that advanced human/machine communication systems must build on our understanding of the human communication process. Despite recent development of some speech recognition systems with high accuracy, the performance of such systems typically falls far short of human capabilities. We are placing heavy emphasis on designing systems that can make use of the knowledge gained over the past four decades on human communication, in the hope that such systems will one day have a performance approaching that of humans.

We are basing the design of our system on the premise that robust speech recognition is tied to our ability to successfully extract the linguistic information from the speech signal and discard those aspects that are extra-linguistic. Like others before us, we have chosen phonemes and other related descriptors such as distinctive features and syllables as the units to relate words in the lexicon to the speech signal. However, there are several aspects that collectively distinguish our approach from those pursued by others. First, we believe that many of the acoustic cues for phonetic contrast are encoded at specific times in the speech signal. Therefore, one must explicitly establish acoustic landmarks in the speech signal in order to fully utilize these acoustic attributes. Second, unlike previous attempts at explicit utilization of speech knowledge by heuristic means, we seek to make use of the available speech knowledge by embedding such knowledge in a formal framework whereby powerful mathematical tools can be utilized to optimize its use. Third, the system must have a stochastic component to deal with the present state of ignorance in our understanding of the human communication process and its inherent variabilities throughout. It is our belief that speech-specific knowledge will enable us to build more sophisticated stochastic models than what is currently being attempted, and to reduce the amount of training data necessary for high performance. Finally, the ultimate goal of our research is the *understanding* of the spoken message, and the subsequent accomplishment of a task based on this understanding. To achieve this goal, we must fully integrate the speech recognition part of the problem with natural language processing so that higher level linguistic constraints can be utilized.

*This research was supported by DARPA under Contract N00039-85-C-0254, monitored through Naval Electronic Systems Command.

This paper describes those parts of our system dealing with acoustic segmentation, phonetic classification, and lexical access, and documents its current performance on the DARPA Resource Management task [1].

SYSTEM DESCRIPTION

There are three major components in the SUMMIT system, as illustrated in Figure 1. The first component transforms the speech signal into an acoustic-phonetic description. The second expands a set of baseform pronunciations into a lexical network. The final component provides natural language constraints. Our preliminary efforts in natural langauge are described in a companion paper [2]. The acoustic-phonetic and lexical components will be discussed in more detail in the following sections.

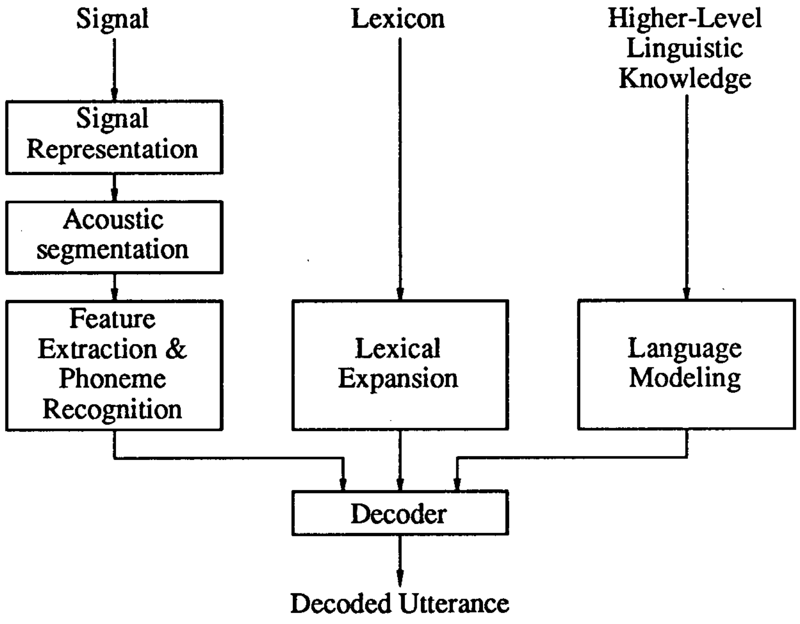


Figure 1: The major components of the SUMMIT system.

ACOUSTIC-PHONETIC REPRESENTATION

The phonetic recognition subsystem of SUMMIT takes as input the speech signal and produces as output a *network* of phonetic labels with scores indicating the system's confidence in the segments and in the accuracy of the labels. The subsystem contains three parts: signal representation, acoustic segmentation, and phonetic classification. In this section, we describe each of these three parts in some detail.

Signal Representation

The phonetic recognition process starts by first transforming the speech signal into a representation based on Seneff's auditory model [3]. The model has three stages. The first stage is a bank of linear filters, equally spaced on a critical-band scale. This is followed by a nonlinear stage that models the transduction process of the hair cells and the nerve synapses. The output of the second stage bifurcates, one branch corresponding to the mean firing rate of an auditory nerve fiber, and the other measuring the synchrony of the signal to the fiber's characteristic frequency.

The outputs from various stages of this model are appropriate for different operations in our subsystem. The nonlinearities of the second stage produce sharper onsets and offsets than are achieved through simple linear filtering. In addition, irrelevant acoustic information is often masked or suppressed. These properties make such a representation well-suited for the detection of acoustic landmarks. The synchrony response, on the other hand, provides enhanced spectral peaks. Since these peaks often correspond to formant frequencies in vowel and sonorant consonant regions, we surmise that the synchrony representation may be particularly useful for performing fine phonetic distinctions. Advantages of using an auditory model for speech recognition have been demonstrated in many contexts, and can be found readily in the literature [4,5,6].

Acoustic Segmentation

Outputs of the auditory model are used to perform acoustic segmentation. The objective of the segmentation procedure is to establish explicit acoustic landmarks that will facilitate subsequent feature extraction and phonetic classification. Since there exists no single level of segmental representation that can adequately describe all the acoustic events of interest, we adopted a multi-level representation that enables us to capture both gradual and abrupt changes in one uniform structure. Once such a structure has been determined, acoustic-phonetic analysis can then be formulated as a path-finding problem in a highly constrained search space.

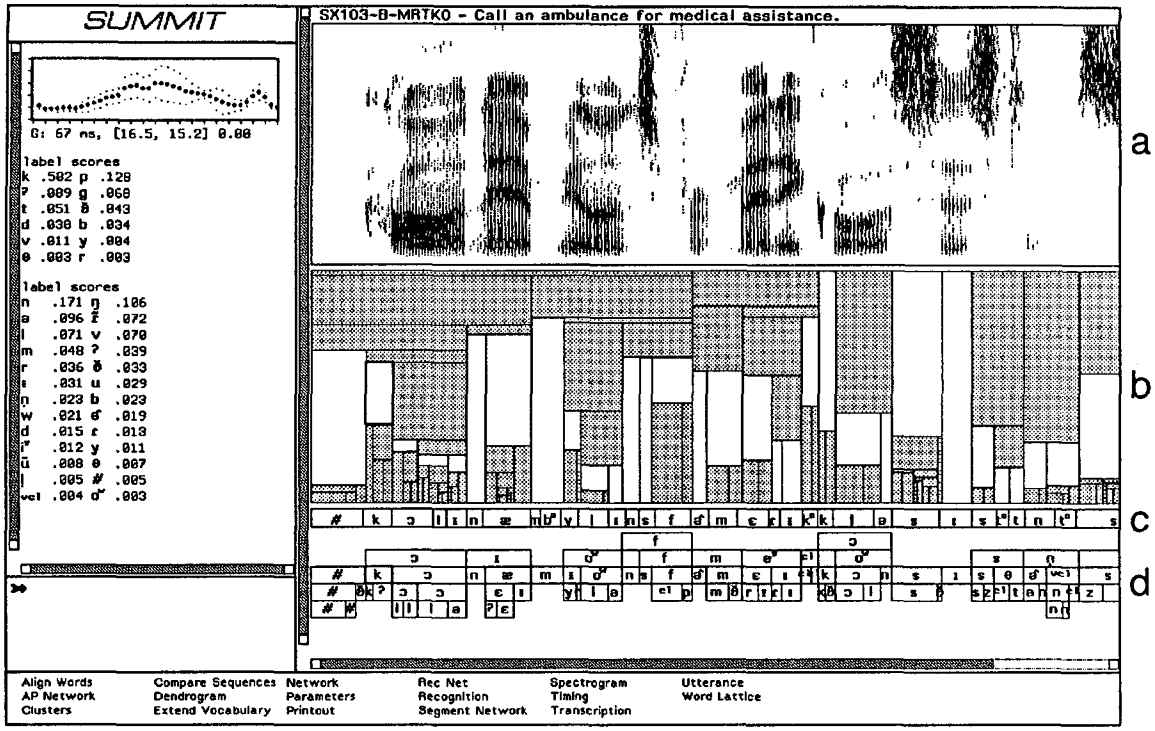
The construction of the multi-level representation has been described elsewhere [7,8]. Briefly, the algorithm delineates the speech signal into regions that are acoustically homogeneous by associating a given *frame* to one of its immediate neighbors. Acoustic boundaries are marked whenever the association direction switches from past to future. The procedure is then repeated by comparing a given acoustic *region* with its neighboring regions. When two adjacent regions associate with each other, they are merged together to form a single region. The process repeats until the entire utterance is described by a single acoustic event. By keeping track of the distance at which two regions merge into one, the multi-level description can be displayed in the form of a dendrogram, as is illustrated in Figure 2 for the utterance "Call an ambulance for medical assistance." From the bottom towards the top of the dendrogram, the acoustic description varies from fine to coarse. The release of the /k/ in "call," for example, may be considered to be a single acoustic event or a combination of two events (release plus aspiration) depending on the level of detail desired. By comparing the dendrogram with the time-aligned phonetic transcription shown below, we see that, for this example, most of the acoustic events of interest have been captured.

Phonetic Recognition

The multi-level acoustic segmentation provides an acoustic description of the signal. Before lexical access can be performed, the acoustic regions must be converted into a form that reflects the way words are represented in the lexicon, which, in our case, is in terms of phonemes. Since some of the phonemes can have more than one stable acoustic region, the mapping between phonemes and acoustic region cannot be one-to-one. Currently, we allow up to two acoustic regions to represent a single phoneme. This is implemented by creating an acoustic-phonetic (AP) network from the dendrogram that includes all single and paired regions. We have experimentally found this choice to be a reasonable compromise between a flexible representation and computational tractability. To account for the fact that certain paths through the AP network are more likely to occur than others, each segment is assigned a weight.

Next, each of the segments in the AP network is described in terms of a set of attributes, which are then transformed into a set of phoneme hypotheses. Rather than defining specific algorithms to measure the acoustic attributes, we define generic property detectors based on our knowledge of acoustic phonetics. These detectors have free parameters that control the details of the measurement. Their optimal settings are established by a search procedure using a large body of training data [11].

This process is illustrated in Figure 3. In this example, we explore the use of the spectral center of gravity as a generic property detector for distinguishing front from back vowels. It has two free parameters, the



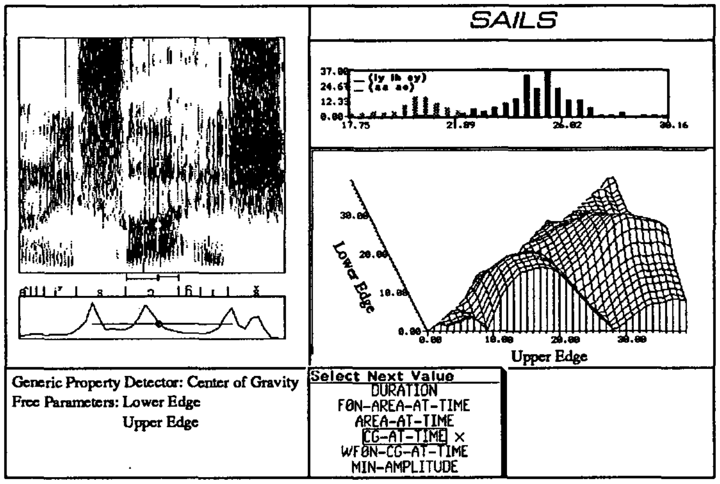
```
G: 67 ms, [16.5, 15.2] 0.00

label scores
k  .502 p  .120
ʔ  .089 g  .060
t  .051 ð  .043
d  .030 b  .034
v  .011 y  .004
θ  .003 r  .003

label scores
n   .171 ŋ  .106
ə   .096 ř  .072
l   .071 v  .070
m   .048 ʔ  .039
r   .036 ð  .033
ɨ   .031 u  .029
ŋ̩   .023 b  .023
w   .021 ɗ  .019
d   .015 ɾ  .013
iʸ  .012 y  .011
ū   .008 θ  .007
l̩   .005 #  .005
vcl .004 dʳ .003
```

Figure 2: Acoustic segmentation of the sentence, "Call an ambulance for medical assistance." The display panel on the right contains: a) spectrogram, b) a dendrogram, c) the time-aligned phonetic transcription, and d) an acoustic-phonetic network.

lower and upper frequency edges. An example of this measurement for a vowel token is superimposed on the spectral slice below the spectrogram, with the horizontal line indicating the frequency range. To determine the optimal settings for the free parameters, we first compute the classification performance on a large set of training data for all combinations of the parameter settings. We then search for the maximum on the surface defined by the classification performance. The parameter settings that correspond to the maximum are chosen to be the optimal settings. For this example, the classification performance of this attribute, using the automatically selected parameter settings, is shown at the top right corner. Note that an attribute can also be used in conjunction with other attributes, or to derive other attributes.

We believe that the procedure described above is an example of successful knowledge engineering in which the human provides the knowledge and intuition, and the machine provides the computational power. Frequently, the settings result in a parameter that agrees with our phonetic intuitions. In this example, the optimal settings for this property detector result in an attribute that closely follows the second formant, which is known to be important for the front/back distinction. Our experience with this procedure suggests that it is able to *discover* important acoustic parameters that signify phonetic contrasts, without resorting to the use of heuristic rules.

Once the attributes have been determined, they are selected through another optimization process. Classification is achieved using conventional pattern classification algorithms [9]. In our current scheme, we use a double-layered approach, with the first layer distinguishing among a small set of classes, and the second layer defining a mapping from these classes to the phone labels used to represent the lexicon. This approach enables us to build a small number of simple classifiers that distinguish the speech sounds along



Generic Property Detector: Center of Gravity
Free Parameters: Lower Edge
Upper Edge

Figure 3: An example of interactive discovery of acoustic attributes for phonetic classification.

several phonetic dimensions. The aggregate of these dimensions describes the contextual variations, which can then be captured in the mapping between the classes and the lexicon. Our experience indicates that such an approach leads to rapid convergence in the models with only a small number of training tokens for each label.

The current scheme for scoring the N classes begins with $N(N-1)/2$ pairwise Gaussian classifiers, each of which uses a subset of the acoustic attributes selected to optimize the discrimination of the pair. The probability of a given class is obtained by summing the probabilities from all the relevant pairwise results. The classes are then mapped to an orthogonal space using principal component analysis. Finally, the score for each phoneme label is obtained from a Gaussian model of the distributions of the scores for the transformed classes.

Following phone classification, each segment in the AP network is represented by a list of phone candidates, with associated probability, as illustrated in Figure 2. The network is shown just below the transcription. In this display, only the AP segments surrounding the most probable path are displayed. The network displays only the top-choice label, although additional information can easily be accessed. For this example, the /k/ in "call" is correctly identified, and its score, in terms of probability, is displayed in the left-hand panel along with several near-miss candidates. On the other hand, the same panel shows that the correct label for the first schwa in "assistance" is the third most likely candidate, behind /n/ and /ŋ/.

LEXICAL REPRESENTATION

We are adopting the point of view that it is preferable to offer several alternative pronunciations for each word in the lexicon, and then to build phoneme models that can be made more specific as a consequence. If accurate pronunciation probabilities can be acquired for the alternate forms, then this is a viable approach for capturing inherent variability in the acceptable pronunciations of words. For example, the last syllable in a word such as 'cushion' could be realized as a single syllabic nasal consonant or as a sequence of a vowel and a nasal consonant. The vowel could be realized as a short schwa, or as a normal lax vowel. For the system to be able to accept all of these alternatives, they must be entered into the lexicon in the form of a network. Currently, lexical pronunciations are expanded by rule to incorporate both within-word and across-word-boundary phonological effects.[1] These rules describes common low-level phonological processes such as flapping, palatalization, and gemination.

We have developed an automatic procedure for establishing probability weighting on all of the arcs in the word pronunciation networks. Currently the weights are entered into the total log probability score and are centered around a score of zero representing no influence. These weights were generated automatically by determining both the recognition path as well as the forced recognition path (i.e., the path obtained when the system is given the correct answer) for a large number of utterances. From this information, we computed: 1) the number of times an arc was used correctly, R, 2) the number of times an arc was missed, M, and 3) the number of times an arc was used incorrectly, W. Once these numbers were tabulated we could assign a weight to each lexical arc. Currently, this weight corresponds to the log ratio of $R+M$, which is the total number of times an arc was used in the forced recognition path, to $R+W$, which is the total number of times an arc was used in the normal recognition path. Thus, if an arc was missed more often than it was used incorrectly, a positive weight is added to the lexical score, which will make the system prefer to use this arc. When the arc is more often incorrect, a negative weight is added, penalizing that arc. When there are the same number of misses as incorrect uses of the arc, or when they form a small fraction of the total number of times an arc was used correctly, the weight has little influence.

DECODER

The lexical representation described above consists of pronunciation networks for the words in the vocabulary. These networks may be combined into a single network that represents all possible sentences by connecting word end nodes with word start nodes that satisfy the inter-word pronunciation constraints. Local grammatical constraints may also be expressed in terms of allowable connections between words.

The task of lexical decoding can be expressed as a search for the best match between a path in this lexical network and a path in the AP network. Currently, we use the Viterbi algorithm to search for this best scoring match. Since we cannot expect the phonetic network to always contain the appropriate phonetic sequence, the search algorithm allows for the insertion and deletion of phonetic segments with penalties that are based on the performance of the AP network on training data. The search algorithm is illustrated in Figure 4.

The possible alignments of nodes in the lexical network to nodes in the phonetic network are represented by a matrix of node-pairs. A match between a path in the lexical network and a path in the phonetic network can be represented as a sequence of allowable links between these node-pairs. The allowable links fall into four categories: normal matches, insertions, deletions, and interword connections. Examples of each are shown in Figure 4. Link (a) is a normal match between an arc in the lexical network and an arc in the phonetic network. Link (b) is an example of an insertion of a phonetic segment (the path advances by a phonetic segment while staying at the same point in the lexical network). Link (c) is an example of an interword connection. Link (d) is an example of a deletion of a phonetic segment (the path contains a lexical arc without advancing in the phonetic network).

The score for a match is the sum of the scores of the links in the match. This allows the search for the best path to proceed recursively since the best score to arrive at a given node-pair is the best of the score of each arriving link plus the best score to arrive at start of the link. Currently, the scores include a phonetic match component, an existence score based on the probability of the particular segmentation, a lexical weight associated with the likelihood of the pronunciation, and a duration score based on the phone duration statistics. The best match for the utterance is the best match that ends at terminal nodes of the lexical network and phonetic network.

[1] Our system currently uses a phonological expansion program, called RULE, developed by researchers at SRI International [12].

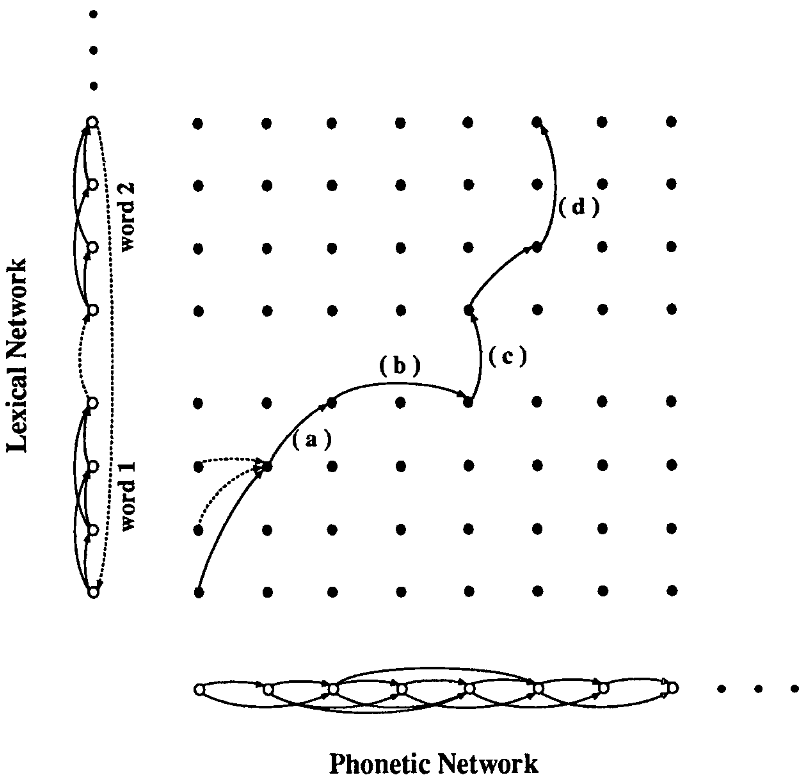


Figure 4: Illustration of Viterbi search used in SUMMIT.

PERFORMANCE EVALUATION

PHONETIC RECOGNITION

The effectiveness of the acoustic-phonetic component has been reported elsewhere [7,13]. The performance of the segmentation algorithm was measured by first finding a path through the dendrogram that corresponds best to a time-aligned phonetic transcription, as illustrated by the path highlighted in white in Figure 2, and then tabulating the differences between these two descriptions. On 500 TIMIT [10] sentences spoken by 100 speakers, the algorithm deleted about 3.5% of the boundaries along the aligned path, while inserting an extra 5%. Analysis of the time difference between the boundaries found and those provided by the transcription shows that more than 70% of the boundaries were within 10 ms of each other, and more than 90% were within 20 ms.

The phonetic classification results are evaluated by comparing the labels provided by the classifier to those in a time-aligned transcription. We have performed the evaluation on two separate databases, as summarized in Table 1. Performance was measured on a set of 38 context-independent phone labels. This particular set was selected because it has been used in other recent evaluations within the DARPA community. For a single speaker, the top-choice classification accuracy was 77%. The correct label is within the top three nearly 95% of the time. For multiple and unknown speakers, the top-choice accuracy is about 70%, and the correct choice is within the top three over 90% of the time. Figure 5 shows the rank order statistics for the speaker-independent case.

Database	No. of Training Sentences	No. of Training Speakers	No. of Test Sentences	No. of Test Speakers	Top-Choice Accuracy (%)
1	510	1	210	same	77
2	1500	300	225	45	70

Table 1: Summary of speaker-dependent and -independent phonetic classification results.

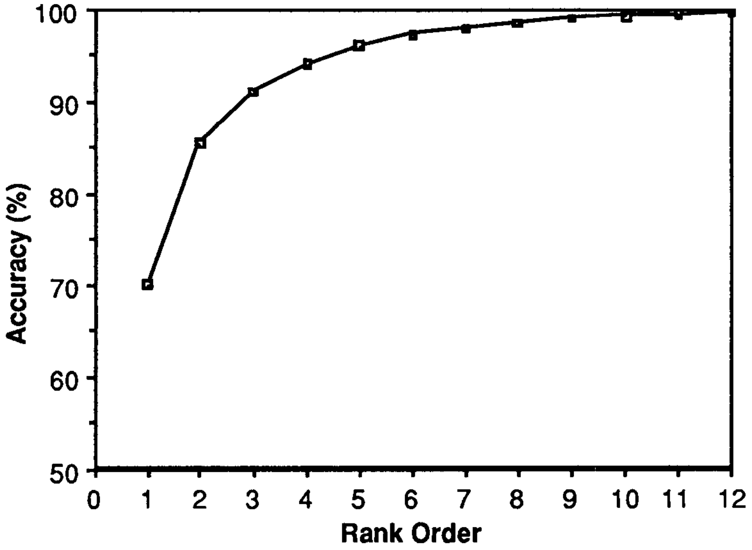


Figure 5: Rank order statistics for the current phone classifier on a speaker-independent task. There are 38 context-independent phone labels: 14 vowels, 3 semivowels, 3 nasals, 8 fricatives, 2 affricates, 6 stops, 1 flap, and one for silence.

WORD RECOGNITION

The SUMMIT system was originally developed for the task of recognizing sentences from the TIMIT database. Over the past three months, we have ported the system to the DARPA 1000-word Resource Management (RM) task, and evaluated its recognition performance. The phoneme models were seeded from 1500 TIMIT sentences, and re-trained on the RM task using 40 sentences each from 72 designated training speakers [1]. The system was evaluated on two test sets, and for two conditions. The first test set, containing 10 sentences each from 15 speakers, is known as the *'87 Test Set.* The second test set, called the *'89 Test Set*, was recently released to the DARPA community, and it contains 30 sentences each from 10 speakers. Each test set was evaluated under both the *all-word* condition (i.e. no language model) and the *word-pair* conditions, in which a designated language model with a perplexity of 60 is used.

The results of our evaluation are summarized in Table 2.[2] Note that this result is obtained by using 75 phoneme models, 32 of which are used to denote 16 stressed/unstressed vowel pairs. At the moment, our system does not explicitly make use of context-dependent models.

[2] The accuracy is computed from the error rate which includes insertions, deletions, and substitutions

Test Set	All-Word Accuracy (%)	Word Pair Accuracy (%)
'87	42.3	87.2
'89	46.2	86.4

Table 2: Summary of word recognition performance results.

DISCUSSION

This paper summarizes the current status of the SUMMIT system. We have described the implementation and reported results for phonetic classification as well as word recognition for the DARPA Resource Management tasks with and without a language model.

Our evaluation results on phonetic classification indicate that performance is much better for a single speaker than for multiple, unknown speakers. This result should not be surprising, since the acoustic variability across speakers is much larger than that within a speaker. One way to assess these results is to compare them to human performance on a similar task. We have conducted some preliminary listening tests in which subjects were asked to identify a phoneme excised from the same database of multiple speakers with minimal contextual information. The results suggest that human performance may be at the 60 to 70% level.

An area where further research is definitely needed is the appropriate representation for acoustic-to-lexical mapping. This includes a more flexible association of phonemes with acoustic segments than is currently allowed, a different choice for the intermediate phonetic representation, and the development of context-dependent models. Presently the choice of the classes in the first layer of the classifier is somewhat arbitrary. We believe that an inventory of classes that is based on distinctive feature theory may be more appropriate. In the first place, the pairwise discrimination analysis is well-suited to a binary feature representation, where phonetic units with contrasting feature values logically define the two sets to be discriminated. Similarly, context-dependent lexical labels in the second stage could also be mapped to a feature-based form to take into account allophonic variations. For example, an /æ/ followed by a nasal and an /æ/ followed by an alveolar would be marked with the right context [+ nasalized] and [+ alveolar], respectively. Thus, the vowel in the word "can" would be pooled with all other instances of /æ/ followed by nasals, and, separately, with all other instances of /æ/ followed by alveolars (/s,t,d/, etc.), to form two distinct second-layer mappings. This approach may provide an elegant way to incorporate context dependency into our recognition system. It may also help to overcome the sparse data problems inherent in very specific context-dependent models.

The RULE system developed by SRI for phonological expansion has been very useful in providing networks of alternate pronunciations. Nevertheless, we have recently initiated an effort to develop a pronunciation expansion program that allows researchers to write phonological rules more efficiently and flexibly, and to conform to the architecture of other parts of the SUMMIT system. We expect that it will be completed within the next two months. The use of lexical weights was found experimentally to improve the performance of the recognition system by a significant amount. Note that a similar procedure could be used incrementally to allow the system to adapt to a particular speaker's pronunciation preferences.

One still-unresolved issue has to do with how to combine the scores for the individual matches to form a total score for lexical decoding. One possibility is to assign equal weight for equal time. Such a scheme results in an inordinately large weight for long sustained vowels as compared with rapid nonstatic sounds such as stop releases. Within the segmental framework, we have the capability to explore several alternative approaches. With a time-normalization scheme, the system can accept a very short erroneous phoneme with a terrible score, but with a per-phoneme weight a reverse effect can occur, where a very *long* badly-matching phoneme can survive because it gets such little weight. We have come up with an approach which essentially accumulates a total score without any normalization, but adds to the log-probability estimate with each update an offset factor that tends to keep the correct answer near zero. This strategy compared favorably with others that we tried, and also required less computation.

The Viterbi search algorithm is a very efficient mechanism for pruning paths when they merge with better-scoring competitors. However, it loses a great deal of its advantage when a true language model capable of natural language understanding is incorporated, because many fewer paths can be collapsed into a single equivalent class. Our hope, however, is that a Viterbi-like pruning strategy can be incorporated into a hierarchical structure representing a syntactic analysis by keeping a record of equivalent subparses locally with each node in the hierarchy. We plan to pursue this kind of strategy when we join our recognizer with a natural language component.

The word recognition performance of SUMMIT is fairly consistent across test sets. While it is always difficult to compare the performance of recognition systems directly, the establishment of standard datasets, language model, and evaluation guidelines has made the task a lot easier. For example, the SPHINX system developed at CMU [14], when evaluated on the *'87 Test Set* using the word-pair grammar, achieved a word recognition rate of 84% and 93% using 48 and 1000 models. Our result of 87% on 75 models is quite competitive, using a very different approach to speech recognition than hidden Markov modelling. However, it is sobering to note that these results fall far short of human performance. For example, we found that the human word recognition rate for the *'89 Test Set* was approximately 99.9% for a single listener. Clearly we still have a long way to go!

Acknowledgments

In developing the SUMMIT system, we have benefitted from our earlier collaboration with researchers from SRI International. In particular, we would like to thank Mitch Weintraub for providing the RULE program, Hy Murveit for writing the first version of the Viterbi decoding algorithm, and Jared Berstein and Mike Cohen for many hours of discussion on lexical representation and expansion. The development of the SUMMIT system has been helped by the programming expertise of Rob Kassel and David Kaufman.

References

[1] Pallett, D., "Benchmark Tests for DARPA Resource Management Database Performance Evaluations," *Proc. ICASSP-89,* May, 1989.

[2] Seneff, S., "TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems," *These Proceedings,* 1989.

[3] Seneff, S., "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing," *Proc. J. of Phonetics,* vol. 16, pp. 55–76, January 1988.

[4] Glass, J. R., and V. W. Zue, "Signal Representation for Acoustic Segmentation," *Proc. First Australian Conference on Speech Science and Technology,* pp. 124–129, November 1986

[5] Hunt, M. J., and C. Lefebvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model," *Proc. ICASSP-88,* pp. 215–218, April 1988.

[6] Cohen, J.R., "Application of an Auditory Model to Speech Recognition," *Proceedings, Montreal Symposium on Speech Recognition*, p. 8, July, 1986.

[7] Glass, J. R., and V. W. Zue, "Multi-Level Acoustic Segmentation of Continuous Speech," *Proc. ICASSP-88,* pp. 429–432, April 1988.

[8] Glass, J. R., "Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition," Ph.D. thesis, Massachusetts Institute of Technology, May 1988.

[9] Duda, R. O., and P. Hart, *Pattern Classification and Scene Analysis,* John Wiley & Sons, Inc., 1973.

[10] Lamel, L. F., R. H.Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop,* Report No. SAIC-86/1546, pp. 100–109, February 1986.

[11] Phillips, M., "Automatic discovery of acoustic measurements for phonetic classification," *J. Acoust. Soc. Am.,* Vol. 84, S216, 1988.

[12] Weintraub, M., and J. Bernstein, "RULE: A System for Constructing Recognition Lexicons," *Proc. DARPA Speech Recognition Workshop,* Report No. SAIC-87/1644, pp. 44–48, February 1987.

[13] Zue, V., J. Glass, M. Phillips, and S. Seneff, "Acoustic Segmentation and Phonetic Classification in the SUMMIT system," *Proc. ICASSP-89,* May, 1989.

[14] Lee, K-F., *Automatic Speech Recognition: The Development of the Sphinx System,* Kluwer Academic Publishers, Boston, 1989.