ADA 221120

ESD-TR-89-280

Technical Report
875

Robust Speech Recognition Using Hidden Markov Models: Overview of a Research Program

C.J. Weinstein
D.B. Paul
R.P. Lippmann

26 February 1990

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEXINGTON, MASSACHUSETTS

Prepared for the Defense Advanced Research Projects Agency under Air Force Contract F19628-90-C-0002.

Approved for public release; distribution is unlimited.

ACCESSION NUMBER
U 352579 A
[illegible] 1990

ARCHIVES
MIT LINCOLN LABORATORY

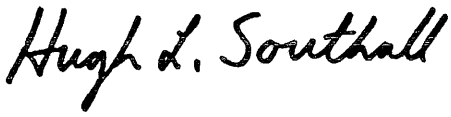
This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. The work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-90-C-0002 (ARPA Order Number 5328).

This report may be reproduced to satisfy needs of U.S. Government agencies.

The ESD Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER



Hugh L. Southall, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission is given to destroy this document when it is no longer needed.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

ROBUST SPEECH RECOGNITION USING HIDDEN MARKOV MODELS: OVERVIEW OF A RESEARCH PROGRAM

C.J. WEINSTEIN
D.B. PAUL
R.P. LIPPMANN
Group 24

TECHNICAL REPORT 875

26 FEBRUARY 1990

Approved for public release; distribution is unlimited.

LEXINGTON MASSACHUSET

MIT LINCOLN LABORATORY
Document Control
ARCH0435467

This page intentionally left blank.

ABSTRACT

This report presents an overview of a program of speech recognition research which was initiated in 1985 with the major goal of developing techniques for robust high performance speech recognition under the stress and noise conditions typical of a military aircraft cockpit. The work on recognition in stress and noise during 1985 and 1986 produced a robust Hidden Markov Model (HMM) isolated-word recognition (IWR) system with 99 percent speaker-dependent accuracy for several difficult stress/noise data bases, and very high performance for normal speech. Robustness techniques which were developed and applied include multi-style training, robust estimation of parameter variances, perceptually-motivated stress-tolerant distance measures, use of time-differential speech parameters, and discriminant analysis. These techniques and others produced more than an order-of-magnitude reduction in isolated-word recognition error rate relative to a baseline HMM system. An important feature of the Lincoln HMM system has been the use of continuous-observation HMM techniques, which provide a good basis for the development of the robustness techniques, and avoid the need for a vector quantizer at the input to the HMM system. Beginning in 1987, the robust HMM system has been extended to continuous speech recognition for both speaker-dependent and speaker-independent tasks. The robust HMM continuous speech recognizer was integrated in real-time with a stressing simulated flight task, which was judged to be very realistic by a number of military pilots. Phrase recognition accuracy on the limited-task-domain (28-word vocabulary) flight task is better than 99.9 percent. Recently, the robust HMM system has been extended to large-vocabulary continuous speech recognition, and has yielded excellent performance in both speaker-dependent and speaker-independent recognition on the DARPA 1000-word vocabulary resource management data base. Current efforts include further improvements to the HMM system, techniques for the integration of speech recognition with natural language processing, and research on integration of neural network techniques with HMM.

This page intentionally left blank.

TABLE OF CONTENTS

This page intentionally left blank.

LIST OF ILLUSTRATIONS

1. INTRODUCTION AND SUMMARY

Since 1985, the Speech Systems Technology Group at Lincoln Laboratory has been carrying out a program of speech recognition research, as part of a multi-laboratory program sponsored by the Defense Advanced Research Projects Agency (DARPA). During the first three years of the effort, the particular focus of the Lincoln program was the development of robust recognition techniques to cope with the stress and noise conditions typical of the fighter cockpit, but which would be applied in situations where limited vocabularies and constrained tasks are acceptable. More recently, the work has moved on to the more general problem of large-vocabulary continuous speech recognition aimed at applications where more natural spoken language input is required.

This report presents an overview of the Lincoln speech recognition program from 1985 through the present. Some highlights of the program are summarized briefly here. The program was initiated in 1985 with the major goal of developing techniques for robust high performance speech recognition under the stress and noise conditions typical of a military aircraft cockpit. The work on recognition in stress and noise during 1985 and 1986 produced a robust Hidden Markov Model (HMM) isolated-word recognition (IWR) system with 99 percent speaker-dependent accuracy for several difficult stress/noise data bases, and very high performance for normal speech. Robustness techniques, which were developed and applied, include multi-style training, robust estimation of parameter variances, perceptually-motivated stress-tolerant distance measures, use of time-differential speech parameters, and discriminant analysis. These techniques and others produced more than an order-of-magnitude reduction in isolated-word recognition error rate relative to a baseline HMM system. An important feature of the Lincoln HMM system has been the use of continuous-observation HMM techniques, which appear to enhance robustness, and avoid the need for a vector quantizer at the input to the HMM system. Beginning in 1987, the robust HMM system has been extended to continuous speech recognition for both speaker-dependent and speaker-independent tasks. The robust HMM continuous speech recognizer was integrated in real-time with a stressing simulated flight task, which was judged to be very realistic by a number of military pilots. Phrase recognition accuracy on the limited-task-domain (28-word vocabulary) flight task is better than 99.9 percent. Recently, the robust HMM system has been extended to large-vocabulary continuous speech recognition, and has yielded excellent performance in both speaker-dependent and speaker-independent recognition on the DARPA 1000-word vocabulary resource management data base. Current efforts include further improvements to the HMM system, techniques for the integration of speech recognition with natural language processing, and research on integration of neural network techniques with HMM, aimed at further improvements in recognition performance.

The organization of this report is as follows. Section 2 describes the stress robustness problem, focussing on the effects of an aircraft environment (e.g., a military fighter cockpit) on speech and on speech recognition. In Section 3 the technical approach to robust recognition is outlined, focussing primarily on robustness enhancements to an HMM isolated-word recognition system. Section 4 summarizes the stress/noise data bases that have been used, and Section 5 outlines the results obtained for isolated-word recognition both for stressed and normal speech. The features of the robust continuous speech recognition system are outlined in Section 6, which actually describes a

class of evolving robust continuous speech recognition systems which have been applied to different conditions and data bases. Section 7 describes the voice-controlled flight simulator, which has been an effective demonstration experiment for real-time speech recognition in a stressful, time-critical task. Section 8 summarizes current results in large-vocabulary continuous speech recognition. New work in the integration of neural networks and HMM is described in Section 9, and Section 10 discusses conclusions and areas for future work.

This report is intended as a summary overview of work at Lincoln performed over several years; the details are covered in the references [1]–[25] published at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) and in other places.

Our work builds on and is influenced by the efforts of numerous past researchers and present colleagues in speech recognition, and particularly in HMM technology. A selection of background HMM and speech recognition references is [26]–[31], which is necessarily only a partial list. The bibliographies of those papers also provide numerous important references. In particular, we have been fortunate to be a part of the multi-laboratory speech recognition program sponsored, since 1985, by the Defense Advanced Research Projects Agency, and have benefited from interactions with our many colleagues in that program. References [32]–[43] provide a representative sampling of speech recognition work at other laboratories participating in the DARPA program from 1985 through the present time.

2. THE PROBLEM OF ROBUSTNESS TO STRESS AND NOISE IN THE AIRCRAFT ENVIRONMENT

A pilot in a high-performance military aircraft operates in a heavy workload environment, where his hands and eyes are busy and speech recognition could be of significant advantage. For example, a speech recognizer could be used to set a radio frequency or to choose a weapon, without requiring hand motion or loss of visual contact with other aircraft or with the terrain. The potential improvement in pilot effectiveness could be extremely significant in critical situations.

A speech recognizer in such a cockpit environment must, however, cope with severe difficulties. The pilot may be exposed to high ambient acoustic noise, encumbered by equipment such as an oxygen mask and headphones, kept busy with a stressful workload task, and subjected to physical and psychological stress. These factors all produce significant changes in the speech signal, which make the recognition task more difficult. Both acoustic noise and headphones (which disturb the auditory feedback path) cause the speaker to talk louder (Lombard effect [44]), with changed spectral tilt and possibly with altered timing (these effects may be somewhat mitigated by the use of sidetone). Generally, it has been found that changes in speech style due to noise are a greater problem than additive noise in the speech, when a facemask with a close-talking microphone is used. Excitement, fear, workload, and distraction produce idiosyncratic speech changes including fast speech, careless speech, or speech block.

In general, acoustic effects due to the stressed environment can include changes in spectral tilt, formant positions, speaking rates, timing, and phonology. Although much study of these effects has been carried out [45],[47], the effects tend to vary idiosyncratically with the speaker and with the situation, and are difficult to predict.

This page intentionally left blank.

3. TECHNICAL APPROACH TO ROBUST RECOGNITION: ROBUSTNESS ENHANCEMENTS TO HIDDEN MARKOV MODEL SYSTEMS

In developing techniques for robust recognition, we chose at the beginning of this research program [1] to build upon the Hidden Markov Model (HMM) approach [26]–[29]. Reasons for choosing this approach included excellent previously-reported recognition performance results in a variety of applications, effective extendability from isolated-word recognition to continuous speech recognition, and (very importantly) trainability from observed data by automatic methods. To our knowledge, we were the first to develop HMM techniques directly focussed on the problem of speaker stress, and the positive results obtained provide strong evidence that HMM provides an excellent framework for developing robust recognition techniques. Currently, HMM-based techniques are widely applied to a large range of speech recognition problems at a variety of laboratories [26]–[29], [32]–[43], [46],[50], and generally have produced the most successful speech recognition systems across both isolated-word and continuous speech recognition tasks.

To form a framework for work in robust recognition, we first developed a baseline HMM system [1,4] using continuous observation HMM techniques, where continuous parameters rather than discrete, vector-quantized symbols (see [28]) were used as input to the recognizer. The use of continuous observations, rather than discrete symbols from a vector quantizer as used in many other current systems [36]–[39], was very important in the development of robustness enhancements, such as the perceptually-motivated distance metric noted below. The training and recognition modules of the baseline isolated-word HMM recognizer are illustrated in Figure 3-1. With reference to the baseline HMM system and word models, note that the term "Hidden Markov Model" refers to the modeling of speech (words, in this case) as a doubly-stochastic process, with an underlying set of states (which can be thought of as states of the speech production mechanism), and Markov transitions between the states. The states are never observed (hence, the term "Hidden"), but for continuous-observation HMM the emissions of the observed parameters in each state (mel cepstra in our models) are modeled according to probability distributions specific to that state. For discrete observation systems, the model includes the probabilities of each vector-quantized symbol being emitted. Once the researcher has specified the form of the model, there is an efficient iterative training algorithm (the Baum-Welch or forward-backward algorithm, see e.g., [28]) which automatically trains the model parameters from training speech (in this case, a number of samples of each word in the vocabulary). For recognition, there is an efficient algorithm (the Viterbi [25]–[29] technique) which chooses the most likely word given a sequence of observation. The baseline system, described in detail in [1,4], uses mel-frequency cepstra [31] computed every 10 ms, as its fundamental observation parameters. The system is termed a "diagonal covariance" system, in that the joint probability density function of the cepstral parameters is assumed to be a multi-variate Gaussian distribution with diagonal covariance matrix. The baseline recognizer uses the standard Baum-Welch iteration for HMM training, and applies a Viterbi recognizer to find the single highest scoring path through the HMM network. The vocabulary word corresponding to this highest score is selected as the recognized word.

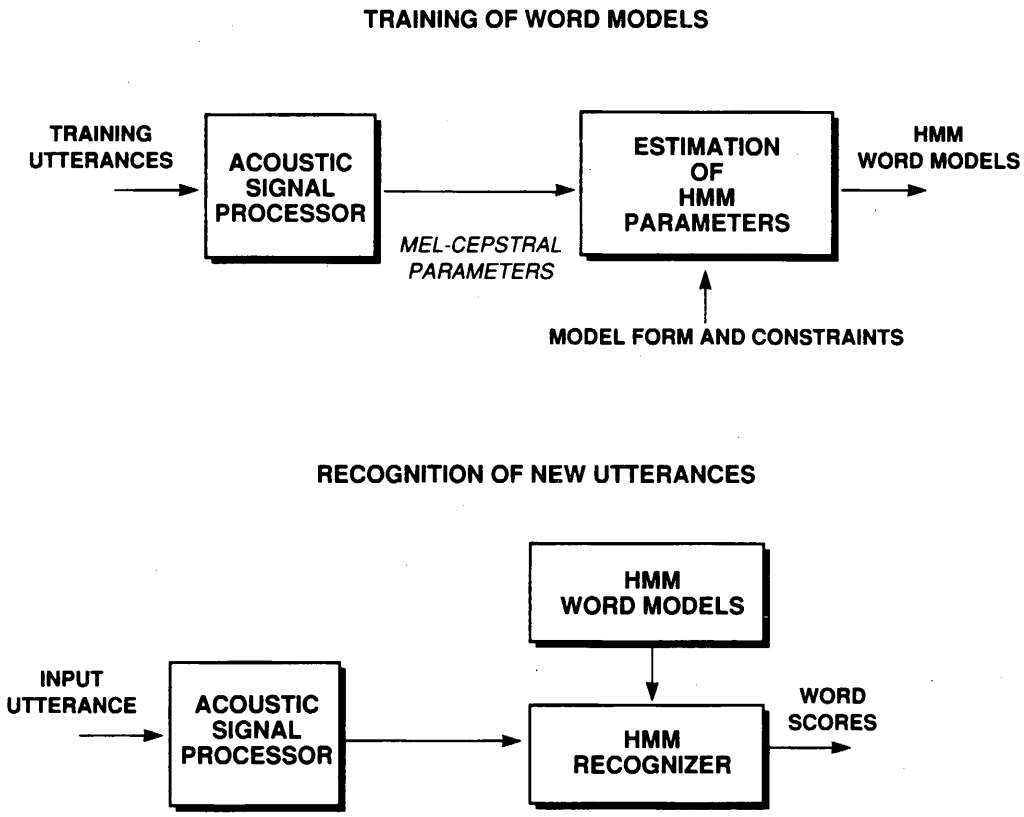


Figure 3-1. Baseline HMM isolated-word recognition system.

Many of the robustness enhancements which have been developed and tested ([1]–[17]) are indicated in Figure 3-2. The specific robustness enhancements, which are numbered and italicized in Figure 3-2 for reference, include enhancements to selection of training utterances [1,7] to the acoustic parameters [1]; to the HMM model form and constraints used in training and recognition [1]–[7]; to background noise modeling [4]; and to discrimination of acoustically-similar words [9,10]. In summary, the key robustness enhancements include:

1. Multi-style training, in which the HMM system is trained on speech spoken in a variety of speech styles to model the variabilities due to stress;

2. Time-differential mel-cepstral parameters, used in addition to the basic mel-cepstral parameters, to better account for the rate of change in the speech signal;

3. Robust parameter estimation techniques, including grand variance (averaging or tying variances over all HMM modes) to compensate for limitations in the amount of training data;

4. A stress-tolerant distance measure using a perceptually-motivated diagonal covariance for HMM nodes, designed to reduce the sensitivity of the recognizer to stress-induced speech variability such as spectral tilt (this distance measure represents an alternative to the robust variance estimation techniques noted in Item 3);

5. Use of improved durational models to better model the time spent in each state of the HMM network (these were found to be very computationally expensive, and not of sufficient benefit to be included in our final systems);

6. Adaptive modeling of the background noise, including facemask breath noise models where appropriate;

7. A cepstral domain stress compensator as a preprocessor for the HMM recognizer;

8. A second-stage, feature-based discriminant analysis system designed to distinguish acoustically-similar words.

Items (1–6) above fit directly into the HMM framework, while (7–8) are integrated into the recognizer, but outside the HMM framework. A variety of other techniques have been developed and tested during the course of this work, including discriminant clustering to reduce the number of parameters and subword models, dynamic adaptation to talker and environment, and a variety of techniques for automatic adjustment of model complexity to the available amount of training data.

The above serves as an outline of the approach to robustness. More discussion on the robustness enhancements, and on their effect on recognition performance, will be presented in the sections to follow.

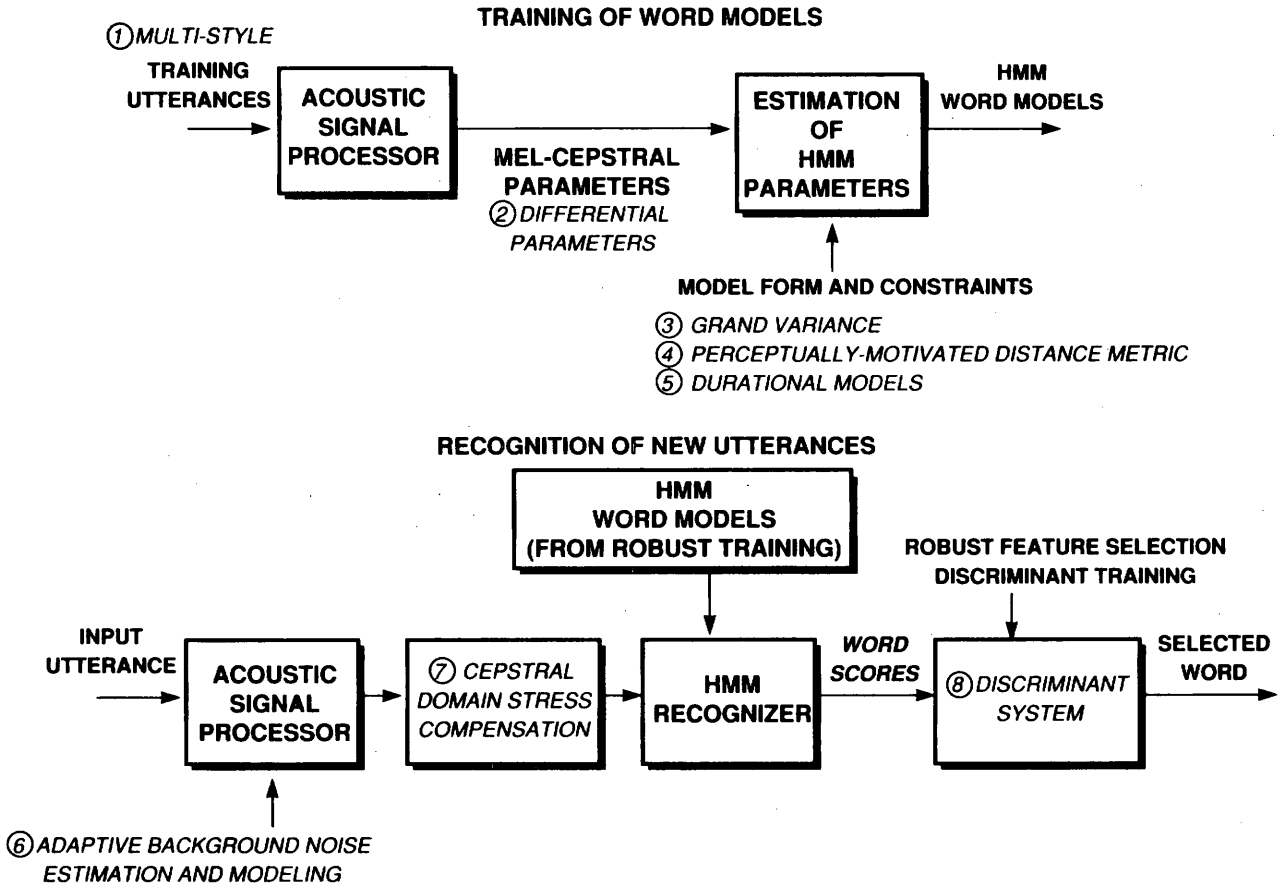


Figure 3-2. Robust HMM isolated-word recognition system with robustness enhancements indicated in italics.

137696-2

4. DATA BASES OF SPEECH PRODUCED UNDER STRESS AND IN NOISE

Collection of a large, systematic data base of speech produced under real stress conditions is a very difficult task. Our approach has been to rely heavily on speech produced with "simulated stress," where the talker is asked to vary the style of speaking to exhibit the range of acoustic variation typical of stressed speech. In addition to style variation, laboratory conditions of workload stress have been utilized. The effects of noise exposure in the ears (Lombard effect [44,45,47]) have been observed to produce similar changes to speech as those produced under stress, and our data bases have generally included the Lombard condition. Two primary data bases have been used for the development and test of robust isolated-word recognition systems.

1. The "TI-stress" data base, a simulated-stress data base provided to us by Texas Instruments [1,42]. This data base uses a 105-word "pilot" vocabulary. It includes 8 talkers (5 male, 3 female) speaking in five talker styles (normal, fast, loud, soft, and shout) and the Lombard condition. There are 5 training utterances (normal speech) per word per talker, and 2 test utterances per word per condition per talker. The data base includes a total of 14, 280 utterances.

2. The "Lincoln-stress" data base [7]-[9], collected at Lincoln. This data base uses a 35-word vocabulary (a subset of the 105-word TI vocabulary) which was selected to include a number of word subsets which are difficult for recognition systems, such as {go,no,oh}, and {six,fix}. It includes 9 talkers (6 male, 3 female) from 3 different dialect groups, with speech produced for 11 conditions: 8 talking styles (normal, slow, fast, soft, loud, clear enunciation, angry, and question pitch); the Lombard condition, and while performing a motor-workload task at two calibrated levels of difficulty. There are 12 training utterances (normal speech) per word per talker, and 2 test utterances per word per condition per talker. The data base includes a total of 10, 710 utterances.

Additional isolated-word recognition experiments have been conducted on a standard, normally-spoken 20-word vocabulary data base [30], collected by Texas Instruments, on which many systems have been tested. This data base is sometimes referred to as "TI-20."

Robust isolated-word recognition experiments and results on all the data bases described above, will be described in Section 5. Some of our robust continuous-speech recognition development and experiments (see Section 7) have been conducted using the "DARPA-robust" continuous speech data base [12,42]. This data base uses a pilot-oriented finite-state grammar and a 207-word vocabulary. Sentences typically range from 4 to 8 words, and the grammar is very constrained. In our work, we used a subset of the data base including 4 training conditions—normal, fast, loud, and 90 dB pink noise, 2 simulated F-16 conditions, and the shout condition. All conditions were recorded using fighter pilot facemasks and headphones. The portion of the data base we used included 5 male speakers, with a total of about 540 sentences per training condition and 220 sentences per test condition.

Recently, most of our continuous speech recognition work has used the normally-spoken, 991-word vocabulary, DARPA resource-management (RM) data base, which is well-documented [41], and has been used by many speech research groups. More description of this data base will be provided in Section 9, where large-vocabulary continuous speech recognition work is described.

5. ROBUST ISOLATED WORD RECOGNITION EXPERIMENTS AND RESULTS

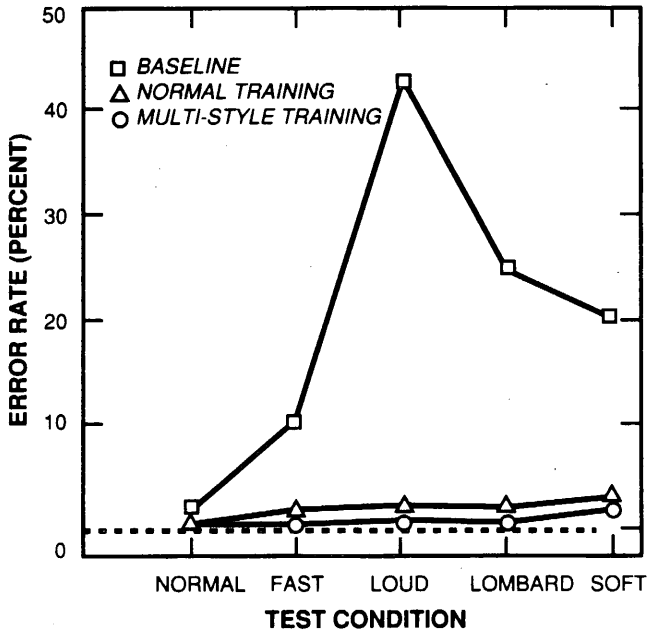
This section summarizes experiments and results in isolated-word recognition of both stressed and normal speech. First, overall results on the TI-stress data base are presented, demonstrating the significant improvements for stressed speech due to the robustness enhancements. The improvements for normal speech are also noted. Then, Lincoln work on a variety of specific enhancement techniques is described, including experiments on both the TI-stress and Lincoln-stress data bases.

Results obtained with the robust HMM isolated-word recognition system are illustrated in Figure 5-1 for the TI-stress data base (105-word vocabulary, 8 talkers, 5 training and 2 test utterances per talker per condition). All experiments were speaker-dependent, in that the system was trained for each speaker using training examples of the vocabulary words spoken by that speaker. Three systems are compared, all of which are diagonal-covariance-matrix, Gaussian probability density systems with mel-cepstrål observations and whole-word models. The "textbook" baseline system uses straightforward training of nodal variances and means. The robust "normal training" system was also trained only on normally-spoken speech, but included enhancements such as perceptually-motivated distance measure, time-differerential parameters, and adaptive background estimation. The robust "multi-style" system included samples of the different speech styles in training (of course the actual test utterances were always separate from the training utterances). Five conditions were tested. The baseline HMM worked reasonably well for normal speech, but degraded for the simulated-stress and Lombard conditions. The improvements with the enhancements are apparent. The percentage substitution rates for normal speech, and for the average over five conditions, are given alongside the figure. For the average over the conditions, more than an order-of-magnitude improvement was achieved. In addition, the robustness enhancements significantly improved performance for normal speech. Best results were obtained using multi-style training. However, the results using the robust system with normal training were also excellent.

The robust isolated-word recognition system, which was developed using the "TI-stress" data base, yielded similar performance results on the "Lincoln-stress" data base, for simulated-stress, Lombard, and workload stress conditions. The system was also tested on the standard TI 20-word vocabulary, normal-speech data base [30] and yielded the best results known to date on that data base—99.94 percent correct, on the first test on that data base. The system also performed well in a number of informal live-input tests, including over long-distance telephone lines. The basic robust HMM isolated-word recognition system, and results and experiments on the TI-stress data base and on the TI 20-word normal speech data base, are described in [4].

The following paragraphs give a summary of Lincoln work on a variety of specific robustness techniques, including experiments on both the TI-stress and Lincoln-stress data bases.

An issue which was investigated early in our work [1] was the relative importance of the following two major effects of noise on recognition in the fighter cockpit: (1) the noise causes the pilot to speak louder and more distinctly (Lombard effect), and (2) the noise leaks into the microphone and degrades the input signal-to-noise (S/N) ratio. Experiments were performed using



PERCENT ERROR RATE

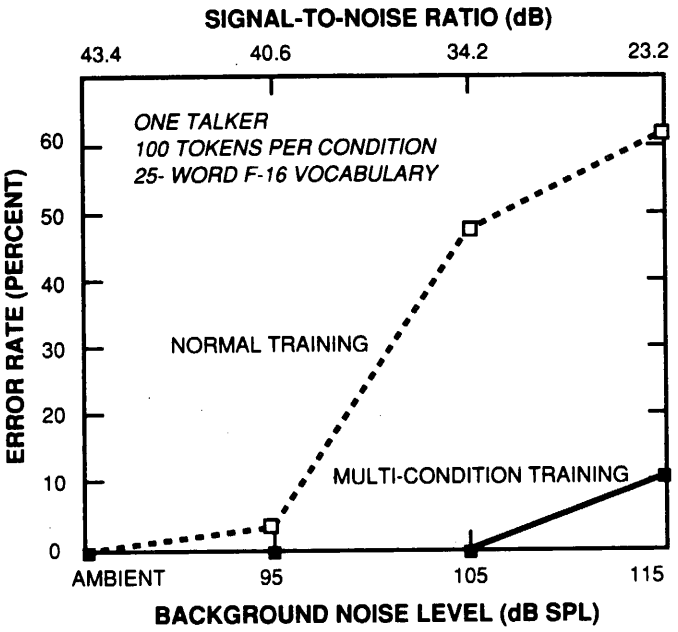
SYSTEM	TEST CONDITION NORMAL	AVG.
BASELINE	1.90	20.49
NORMAL TRAINING	.24	1.86
MULTI-STYLE TRAINING	.00	.71

137696-3

Figure 5-1. Performance of robust HMM-word recognition system on the 105 word vocabulary TI-stress data base. All experiments are speaker-dependent. The substantial improvements over the baseline system for the simulated-stress and noise exposure (Lombard) conditions are apparent. The robustness enhancements also improved the performance for normal speech.

recordings made at the Wright-Patterson Air Force Medical Research Laboratory. Words in a 25-word vocabulary were produced by one talker wearing a facemask and helmet in an ambient condition and with simulated fighter aircraft (F-16) background noise levels of 95, 105, and 115 dB sound pressure level (SPL). The S/N ratio at the recognizer input (after the noise-cancelling microphone) was determined for the ambient training condition, and for training samples using speech collected under the multiple noise conditions. Recognition results are shown in Figure 5-2. Recognition results for normal training indicate severe performance degradation in noise, although the S/N ratio remains high (23 dB) even for the highest noise level. However, good performance is achieved with multi-condition training using training data obtained under multiple noise conditions. These data, together with careful listening to the recordings, strongly suggest that the degraded performance was due to the Lombard effect, and not due to additive noise. These results, which are consistent with those presented in [42], indicate that the Lombard effect is more important than

additive noise, at least in an environment where a noise-cancelling microphone is used. The results in Figure 5-2 also demonstrate that training under multiple conditions is an effective technique to compensate for the Lombard effect.



137696-4

Figure 5-2. HMM system performance in a simulated F-16 noise environment with normal training and with training under multiple noise conditions.

The focus of [5,6] is a particular robustness enhancement technique wherein the basic recognition parameters (mel-frequency cepstra) are modified adaptively to compensate for variations due to stress. This adaptation is shown to compensate for spectral tilt and to produce significant performance improvements for isolated-word recognition systems trained with normal speech. The studies conducted in developing these compensation techniques also yielded important data [5,6] on the specific effects of various speaking styles on the cepstral parameters.

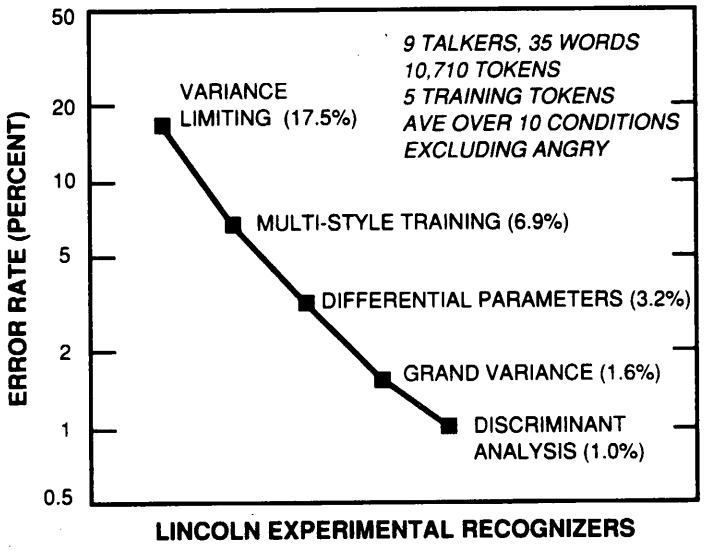
Multi-style training, and experiments and results on the Lincoln-stress data base, are the focus of [7,8]. The effectiveness of training on multiple talker styles in improving recognition performance for stress and noise conditions (workload, Lombard) not included in the training data is reported and discussed. Overall recognition accuracy of 99 percent on the difficult Lincoln-stress data is reported, achieved via a combination of multi-style training and other robustness enhancements.

A second-stage discriminant-analysis system, developed to serve as a post-processor to the HMM recognizer in order to resolve confusion between acoustically-similar words, is described in [9, 10]. This discriminant system is trained by passing samples of every word in the vocabulary through the HMM models of every word in the vocabulary, to explicitly model acoustic differences between words. A statistically-based sifting technique is described which selects only those parameters which are likely to be effective in discrimination. Performance improvements relative to the robust single-stage HMM are reported for the Lincoln-stress data base, contributing, for example, to the overall 99 percent recognition accuracy on that data base.

An illustrative summary [7] of the effects of various robustness techniques on performance with the Lincoln-stress data base is shown in Figure 5-3. The basic HMM system, augmented with a variance-limiting technique to prevent underestimates of parameter variances, achieved 17.5 percent error rate (averaged over 9 talkers and 10 conditions). Multi-style training reduced the error rate to 6.9 percent, and the use of differential cepstral parameters (in addition to the basic cepstral parameters) reduced error rate further to 3.2 percent. Grand variance techniques, which reduce the effect of limited training data by estimating cepstral parameter variances as an average over all words and phonemes (the means are separately estimated), lowered the error rate to 1.6 percent. Finally, the second-stage discriminant analysis corrected enough of the remaining confusion to reduce the error rate on this data base to 1 percent.

A technique for dynamic adaptation of HMM isolated-word model parameters to new speakers and to stress-induced speech variations is described in [13]. Tests were performed on the Lincoln-stress data base. Results of these speaker adaptation experiments are illustrated in Figure 5-4. It was found to be crucial to allow user feedback to assist the adaptation process by correcting errors produced by the system. This is illustrated in Figure 5-4, by the difference between the case of adaptation on all tokens (the user must supply the correct answer when an error is made) and the case of adaptation only on correct recognition. With corrective feedback from the user provided, speaker-adaptation experiments produced error rates equivalent to speaker-trained systems after presentation of only a single new token per vocabulary word. Stress-condition adaptation experiments produced results comparable to multistyle-trained systems after the presentation of several new tokens per vocabulary word. Similar adaptation techniques, focusing primarily on adaptation to noise conditions, are presented in [51].

Finally, a training procedure called discriminant clustering for automatically generating subword HMM models for an IWR system is presented in [14]. (A similar technique for template-based recognition is described in [46]). HMM node sequences from whole-word models were merged using statistical clustering techniques. This procedure reduced computation during recognition (on the Lincoln-stress data base) by roughly one-third without significant increase in error rate. Additional clustering work, focusing specifically on triphones, is described in [12].



137696-5

Figure 5-3. Effects of robustness techniques on HMM recognition performance on the Lincoln-stress 35-word vocabulary data base.

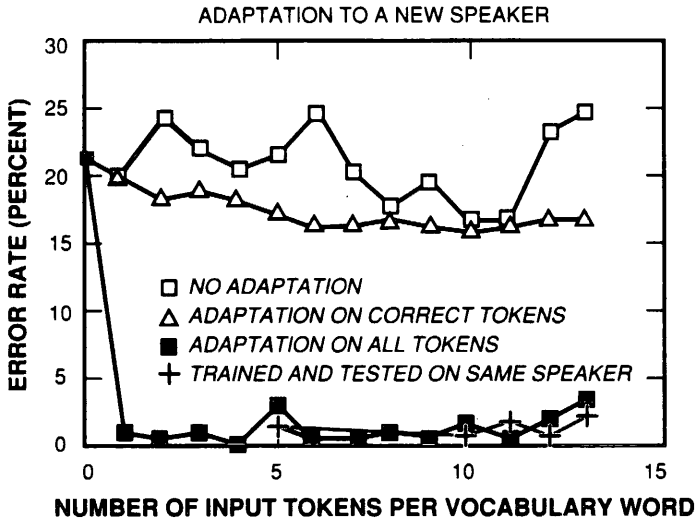


Figure 5-4. Experiments on adaptation of HMM recognizer to a new speaker (normal speech). All points are averaged over 9 speaker pairs where initial training models are obtained from one speaker then tested on another.

6. ROBUST CONTINUOUS SPEECH RECOGNITION SYSTEM

The performance improvements achieved on simulated-stress isolated-word data bases led us to extend the robust HMM work to continuous speech recognition. Although isolated-word recognition could be useful in environments such as the cockpit, users generally would prefer continuous speech to isolated words, as long as high recognition accuracy is maintained. A fairly restrictive grammar of command phrases would be useful to pilots and could constrain the recognition task sufficiently to provide good performance in the cockpit environment.

A key step in moving to a continuous speech recognition system was to change from whole-word to subword (phone) models. In particular, this removed the need to train on all words in the vocabulary. In deriving subword models, we first derived from a dictionary a representation of each word as a sequence of phones. Phone context was taken into account by allowing separate context-dependent phone models, referred to as triphones ([12,36,38]), for each left and right context in which a phone occurred. Phones (or triphones) were modeled as linear sequences of HMM nodes, and words were modeled as linear sequences of triphones.

For the continuous speech recognition system it was also necessary to incorporate word order constraints (syntax and semantics). A finite-state grammar was introduced, which could be adapted to a variety of tasks of different difficulty. The primary measure of difficulty for a grammar in this work is perplexity, defined as two raised to the power of the entropy of the language, or the geometric average of the branching factor (number of words allowed to follow a given word in the grammar).

Robustness features which were extended from the isolated-word to the continuous speech system included the basic approach of continuous observation HMM, mel-cepstral observations with temporal differences, perceptually-motivated distance measures or tied variances, and adaptive background models.

Some of the additional features that have been developed and tested in various versions of the robust continuous speech recognition system include:

1. Training of triphone models using an unsupervised monophone bootstrap so that only an orthographic transcription of the training data (no hand-marking) is needed;

2. Gaussian mixtures of variable order to model the probability density functions of the observations as weighted sums of Gaussian densities, rather than as a single Gaussian for each probability density function;

3. Word-context-dependent triphone models to account for interword context;

4. Extrapolation to model missing triphones.

More detail on the development, evolution, and testing of the robust CSR system is given in [12,15,16,17]. A sampling of results is presented in the next three sections.

This page intentionally left blank.

7. CONTINUOUS SPEECH RECOGNITION EXPERIMENTS ON THE DARPA-ROBUST DATA BASE

The continuous speech recognition system was first developed and tested on the "DARPA-robust" data base [12] (see Section 4). Performance on a 207-word, perplexity-14 task was 2.5 percent word error rate (best speaker) and 5 percent (4-speaker average) for the normal condition of the data base. Performance was poorer under the various conditions of simulated stress and noise, which indicated that this was a rather difficult data base. However, there were a number of problems in gathering this data base, including poor speaker motivation and equipment problems, which may have unduly increased the difficulty of the recognition task. These problems unfortunately may have masked the effects we were trying to observe, namely, those due to stress and noise.

A small data base using the same vocabulary and grammar was gathered for a motivated speaker under office conditions, and a word error rate of 0.9 percent was achieved. This indicated that the vocabulary and grammar were not particularly difficult, and that the problems were in the speech recordings themselves. Our observations and results on the DARPA-robust data base were similar to those of our colleagues at Texas Instruments [43]. Based on the general problems observed with the DARPA-robust data base, we elected to further develop and test the continuous speech recognition system on other data bases, primarily the DARPA Resource Management data base.

This page intentionally left blank.

8. VOICE-CONTROLLED FLIGHT SIMULATOR: A DEMONSTRATION OF REAL-TIME CONTINUOUS SPEECH RECOGNITION UNDER STRESS

In order to demonstrate and test the capabilities of our robust recognition system in a real-time task which simulates the type of stress typical of the cockpit environment, we have developed a voice-controlled flight simulator as illustrated in Figure 8-1. The task is takeoff, flight, and landing of a voice-controlled, simulated F-15 aircraft. It is not intended to model an actual flight task, as we do not recommend that critical functions, such as landing, be handled by voice. However, the real-time interactive environment was intended to be typical of the flight environment, where pilots might control noncritical tasks by voice (see Section 2). The F-15 flight simulator was developed at Lincoln (by D. Paul) and was judged to be realistic by a number of military pilots. Figure 8-1 shows the speech recognizer, a depiction of a typical flight pattern, and the flight simulator display as viewed by the operator. The flight simulator, as well as the HMM recognizer, operate in real-time on a SUN-4 computer. The signal processing front-end is implemented in real-time on a Lincoln-built signal processing computer.

Although the demonstrations are conducted in a relatively benign acoustic environment, the speaker stress is real, due largely to the time-critical nature of the task. The system uses the HMM continuous speech recognizer with a 28-word vocabulary, perplexity-7 grammar to control the aircraft. In one series of demonstrations requiring over 1000 command sentences, only one error occurred. Our general experience is that the correct phrase is recognized more than 99.9 percent of the time. The results for this system demonstrate that speaker stress for a motivated speaker can be tolerated for a simple recognition task, even when very high recognition accuracy is required.

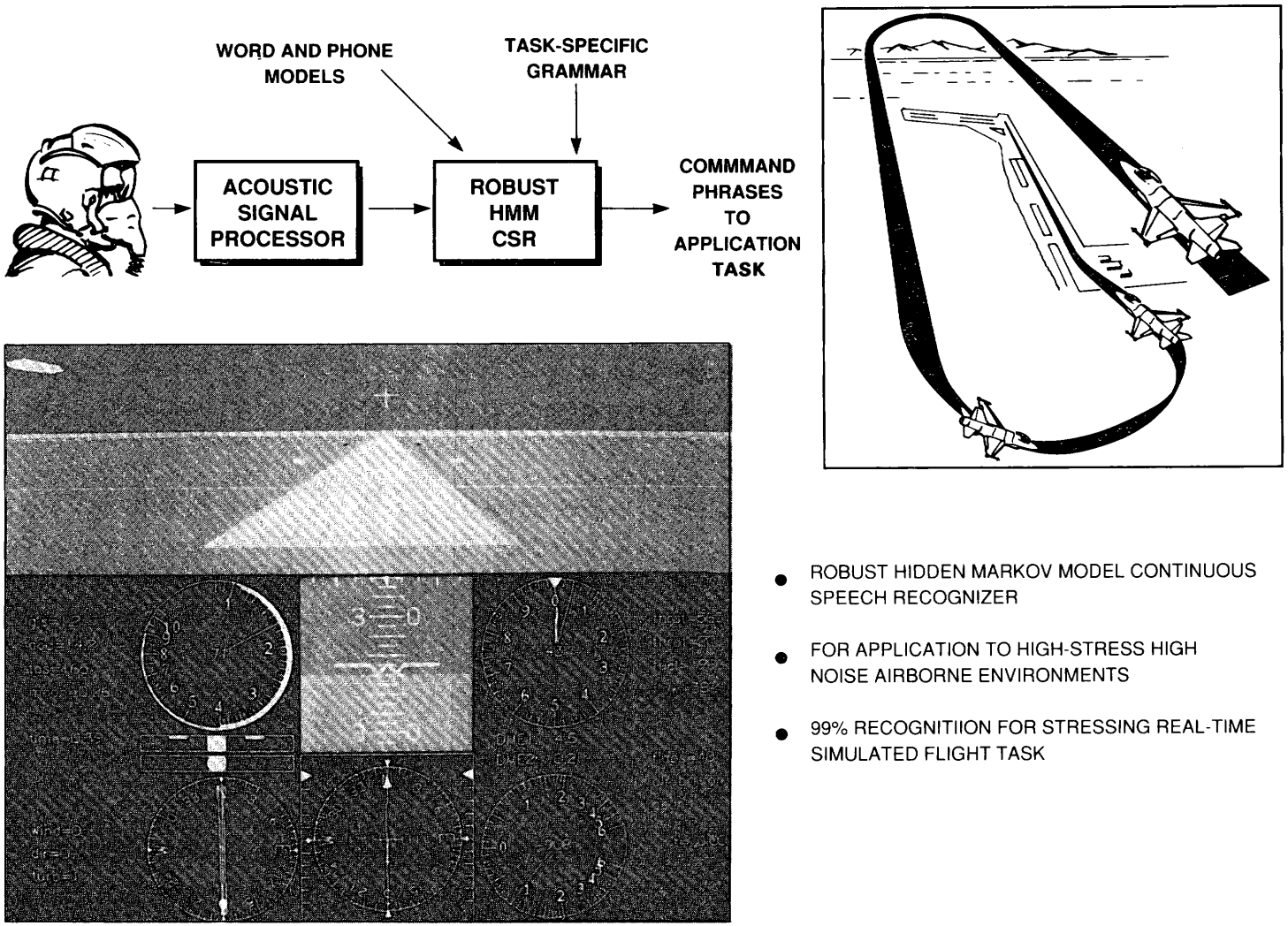


Figure 8-1. Voice-controlled flight simulator, which demonstrates real-time continuous speech recognition under task-induced stress. Illustrated are the speech recognizer, a typical flight pattern, and the flight simulator display.

137696-7

9. LARGE-VOCABULARY CONTINUOUS SPEECH RECOGNITION

Recently, the Lincoln stress-resistant HMM CSR has been extended to large-vocabulary, normally-spoken, continuous speech recognition for both speaker-dependent (SD) and speaker-independent (SI) tasks. Development and test for this effort have been carried out using the DARPA Resource Management data base [41] which is being widely used by a number of other organizations. This data base consists of sentences typical of those which would be spoken in a Naval resource management task, allowing data retrieval and management of ships and other Naval resources. The speech is normally-spoken, the vocabulary is 991 words, and the average sentence length is eight words. Tests were run both with an "official" word-pair grammar (list of allowable word pairs, no assigned probabilities) with a recognition perplexity of 60 and with "no grammar" (all word pairs allowed), corresponding to a perplexity of 991. The speaker-dependent portion of the data base has 12 speakers, with 600 training sentences per speaker and 100 development test sentences per speaker. For speaker-independent work, we trained on 2880 sentences from 72 speakers (SI-72) or 3990 sentences from 109 speakers (SI-109), and used the same development test data as for speaker-dependent. In the DARPA program, a series of official evaluation tests (October 1987, June 1988, February 1989, October 1989) have been run on new data not used in either training or development.

The general features of the Lincoln large-vocabulary CSR system are as summarized in Section 6. The details of the system have continued to be refined, with the goal of improved performance, in response to experimental results obtained during development tests and in official evaluation tests. Throughout this work, the Lincoln continuous-observation HMM continuous speech recognizer has achieved performance on the DARPA resource management task similar to that of the other leading DARPA research groups, all of which use discrete observation HMM. In addition, the Lincoln system has been the only one to be tested in both speaker-dependent and speaker-independent tests for all the official evaluations.

The DARPA resource management tests have all been for normal speech, but the earlier Lincoln work provides evidence that continuous observation HMM has robustness advantages for stressed speech. Current work on other problems requiring robustness (e.g., spotting of key words in continuous telephone speech) has also focused on continuous observation HMM.

To illustrate some highlights of the Lincoln work in large-vocabulary continuous speech recognition, it is useful to summarize improvements made in the system between the June 1988 and February 1989 official tests [15,16]. These include word-context-dependent triphone models, variable mixtures, and tied mixtures [16,17] (Note: tied mixtures were developed, but not actually used in the official February 1989 tests). Work on tied mixtures at other laboratories is described in [52,53].

Word-context-dependent models extend the triphone context to include the phone on the other side of a word boundary. For example, the "ee" in "three words" would have a triphone model distinct from the "ee" phone in "three phones." Word-context-free models would not distinguish

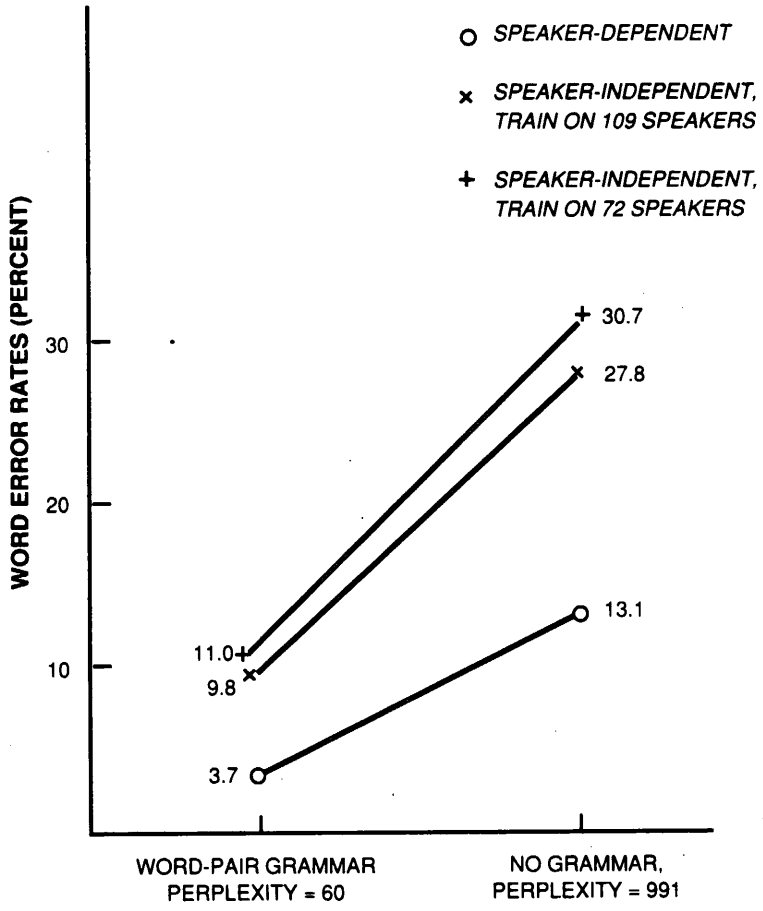
between those two cases. Training and recognition strategies were developed for word-context-dependent models. For training, the training data is used twice per iteration of the Baum-Welch training algorithm—once to train word-context-free triphone models and once to train observed word-context dependent triphones. For recognition, a significant increase in complexity is needed to account for the various word-boundary topologies. The speaker-dependent system using word-context-dependent triphones achieved significant improvements over the word-context-free system—3.0 percent versus 5.2 percent error rate on development data. However, the speaker-independent word-context-dependent results were worse than for word-context-free triphones, indicating that the word-context-dependent system may be too detailed a model for the available speaker-independent training data.

Variable-order Gaussian mixtures were introduced to attempt to match the complexity of the model (the number of Gaussians per mixture) to the available amount of training data (the number of occurrences of a particular triphone in the training data). Small improvements were obtained for the speaker-independent task, and the results indicated that the basic idea was successful but that the function chosen to select the mixture order was not optimum.

Finally, a version of tied mixtures was tested and shown to provide a small improvement for the speaker-independent task. This technique shares Gaussians among different phones (different weights are used for each phone) and, hence, reduces the numbers of Gaussians. It provides another form of matching model complexity to the amount of training data by allowing the system to automatically reduce the number of degrees of freedom in the model when there is insufficient training data.

As a sample of results obtained with the Lincoln continuous speech recognition system on the DARPA Resource Management data base, Figure 9-1 presents results obtained on the February 1989 evaluation test set. A full set of tests was run for both speaker-dependent and speaker-independent recognition, and results are similar to those of the other high performing DARPA-supported groups. The word error results, with grammar, of 3.7 percent (speaker-dependent) and 9.8 percent (speaker-independent), represent state-of-the-art performance, but performance needs to be improved substantially to achieve acceptable sentence accuracy. More work in basic speech recognition, as well as in application of syntax and semantics, is needed and is in progress at Lincoln and at numerous other laboratories.

The results in Figure 9-1 illustrate the substantial impact of the grammar perplexity, as well as the comparison between speaker-dependent and speaker-independent training. In addition, comparisons are provided of speaker-independent performance of the same system trained both on 72 speakers (SI-72) and on 109 speakers (SI-109), with better results occurring for the better-trained system. It is probable that error rate could be further reduced by simply increasing the amount of training data.



137696-8

Figure 9-1. Word error rates for the Lincoln HMM recognizers on the February 1989 "official tests" on the DARPA Resource Management data base.

This page intentionally left blank.

10. INTEGRATION OF NEURAL NETWORKS AND HIDDEN MARKOV MODELS

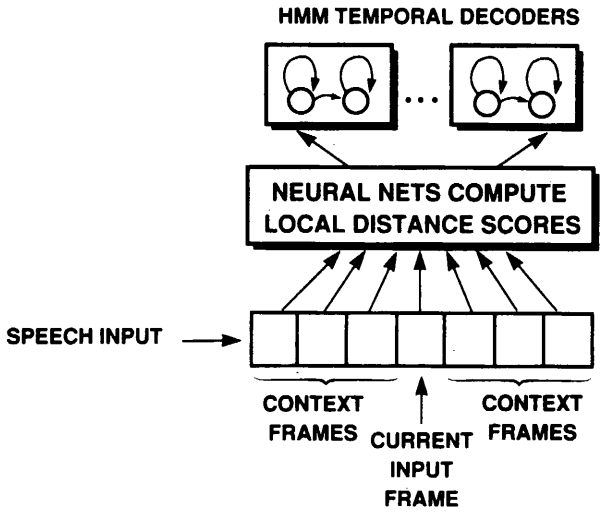
An important area being explored for the improvement of speech recognition performance is the development and application of neural network classification algorithms and architectures. Neural nets offer the potential of new algorithms, dynamic adaptation, and computational efficiency, and have been the subject of intensive efforts at a number of laboratories. Work at Lincoln in neural nets for speech recognition began in 1987, with a major focus on comparison of neural net and conventional pattern classification algorithms [20]–[23], and on efficient neural net implementations of conventional recognizers, including a neural net implementation of the Viterbi decoder used in HMM [23].

Speech recognition using neural networks has so far led to good results only for small-vocabulary tasks that involve low-level units of speech, such as phonemes and letters in the "E-set" [20]. Such problems take advantage of the relatively well-developed status of static neural net classifiers, and of their capability for discrimination between similar patterns. However, research on neural net dynamic pattern classifiers, which are required for more difficult isolated-word and continuous speech tasks, is only at a beginning stage and has met with limited success.

Our current approach is to take advantage of both neural network (NN) classifiers and the HMM framework (which handles the time dimension in dynamic pattern classification) by combining the two techniques. Two approaches [25] are being pursued in developing combined HMM/NN recognizers. One approach applies a multi-layer perception (MLP) as a second-stage discriminator designed to overcome HMM's weakness in focusing on segments of the input speech that are important for discrimination. The motivation of this two-stage scheme is similar to earlier discrimination work described in [9,10], but takes advantage of the automatic discrimination training capability of the back-propagation algorithm. This two-stage HMM/NN approach has yielded improved recognition accuracy for small vocabulary ("B," "D," "G") tasks, but scaling problems have been encountered for larger vocabularies. Various approaches are being investigated for overcoming these problems with larger vocabularies.

A second integration strategy, which is a primary focus of our current work, is to use a neural net for acoustic/phonetic feature extraction to provide local distance scores as shown in Figure 10-1 [25]. Other workers [48,49] have obtained promising results using similar approaches. Our approach is distinguished from others in that we are using fully-automated training, without need for hand-segmenting training data. The training algorithms being explored integrate back-propagation and other neural network algorithms into the iterative forward-backward training algorithm. Although training is very computationally intensive, recognition is not much more complex than a standard HMM system. The promise of the hybrid system shown in Figure 10-1 is that it uses the neural network to extract features while using the HMM for time alignment. Currently, we have implemented a basic hybrid system and have obtained initial E-set results which are similar to existing high performance HMM systems. However, we have not yet exploited frame context, as illustrated

in Figure 10-1, and also are introducing a number of refinements into the training and recognition algorithms.



137696-9

Figure 10-1. System framework for integration of neural network classifiers with HMM recognition, where the neural net is used for acoustic-phonetic feature extraction over multiple input speech frames, and the HMM is used for time-alignment and temporal decoding.

11. CONCLUSIONS AND FUTURE WORK

The work described in this report began with a focus on robust isolated-word recognition in stressed, noisy environments typical of the aircraft environment. Enhancements to a baseline HMM system were developed which significantly improved recognition performance under difficult conditions and which also improved performance for normal speech. Later, the robust HMM system was extended successfully to continuous speech recognition under stress, and then to large-vocabulary continuous speech recognition for normal speech. In addition to the HMM work, Lincoln has successfully developed neural net algorithms and architectures for speech recognition, and current work includes hybrid HMM/NN systems.

Current and projected future areas of Lincoln research in speech recognition include (1) development of robust techniques for talker-independent recognition and key word spotting [19] on noisy and distorted speech; (2) research into the application of speaker recognition techniques to improve speech recognition performance; (3) continued development and improvement of HMM-based continuous speech recognition techniques, including tied mixtures and extensions beyond the basic HMM structure; (4) development of structures, including a new interface specification based on a stack controller [18], for integrating speech recognition and natural language systems into spoken language systems; and (5) continued research on neural network techniques, including improved neural network classifiers and hybrid HMM/NN systems.

This page intentionally left blank.

ACKNOWLEDGEMENTS

We would like to acknowledge the technical contributions of our colleagues Yeunung Chen, Edward.Martin, and Molly Mack to the work on robust HMM speech recognition. Their specific contributions are cited in the references.

We would like to acknowledge the continuing support and encouragement of Dr. Allen Sears, the DARPA Program manager for this work from 1985 to 1988; his insights regarded the aircraft environment and his suggestion to develop a realistic flight-oriented demonstration system were particularly helpful. We would also like to acknowledge the continued support, encouragement, and useful suggestions of Mr. Charles Wayne, who has been the DARPA Program manager for this work since Fall 1988.

The neural network effort at Lincoln was supported under our ongoing Air Force program, first by an internal Lincoln research grant, and later under outside DoD funding. Work is currently continuing under DARPA and AFOSR sponsorship.

This page intentionally left blank.

REFERENCES

1. D.B. Paul, R.P. Lippmann, Y. Chen, and C.J. Weinstein, "Robust HMM-Based Techniques for Recognition of Speech Produced Under Stress and in Noise," in *Proc. DARPA Speech Recognition Workshop*, Palo Alto, CA (1986), pp. 81-92;

   also in *Proc. Speech Tech '86*, New York, NY (1986), pp. 241-249.

2. D.B. Paul, "Training of HMM Recognizing by Simulated Annealing," in *Proc. ICASSP '85*, Tampa, FL (1985), pp. 13-16.

3. C.J. Weinstein, "Robust Speech Recognition: An Update," in *Proc. DARPA Speech Recognition Workshop*, San Diego, CA (1987), pp. 82-84.

4. D.B. Paul, "A Speaker-Stress Resistant HMM Isolated Word Recognizer," in *Proc. ICASSP '87*, Dallas, TX (1987), pp.713-716 also in *Proc. DARPA Speech Recognition Workshop*, San Diego, CA (1987), pp. 85-89.

5. Y. Chen, "Cepstral Domain Stress Compensation for Robust Speech Recognition," in *Proc. ICASSP '87*, Dallas, TX (1987), pp. 717-720; also in *Proc. DARPA Speech Recognition Workshop*, San Diego, CA (1987), pp. 90-95.

6. Y. Chen, "Cepstral Domain Talker Stress Compensation for Robust Speech Recognition," Lincoln Laboratory, Lexington, MA, Technical Rep. 753 (1986). DTIC AD-A176068.

7. R.P. Lippmann, E.A. Martin, and D.B. Paul, "Multi-Style Training for Robust Isolated-Word Speech Recognition," in *Proc. ICASSP '87*, Dallas, TX (1987), pp. 705-708; also in *Proc. DARPA Speech Recognition Workshop*, San Diego, CA (1987), pp. 96-99.

8. R.P. Lippmann, M.A. Mack, and D.B. Paul, "Multi-Style Training for Robust Speech Recognition Under Stress," *J. Acoust, Soc, Am.*, Supplement 1, 79, 595 (1986).

9. E.A. Martin, R.P. Lippmann, D.B. Paul, "Two-Stage Discriminant Analysis for Improved Isolated-Word Recognition," in *Proc. ICASSP '87*, Dallas, TX (1987), pp. 709-712; in *Proc. DARPA Speech Recognition Workshop*, San Diego, CA (1987), pp. 100-104.

10. E.A. Martin, "A Two-Stage Isolated-Word Recognition System Using Discriminant Analysis," MIT Lincoln Laboratory, Lexington MA, Technical Rep. 773 (1987). DTIC AD-A187425.

11. D.B. Paul, "Robust Speech Recognition for Stressful Airborne Environments," in *Proc. Military Speech Technology '87, Media Dimensions*, Arlington, VA (1987), pp. 153-155.

12. D.B. Paul and E.A. Martin, "Speaker Stress-Resistant Continuous Speech Recognition," in *Proc. ICASSP '88*, New York NY (1988), pp. 283-286.

13. E.A. Martin, R.P. Lippmann, D.B. Paul, "Dynamic Adaptation of Hidden Markov Models for Robust Isolated-Word Speech Recognition," in *Proc. ICASSP '88*, New York, NY (1988), pp. 52-55.

14. R.P. Lippmann and E.A. Martin, "Discriminant Clustering Using an HMM Isolated-Word Recognizer," in *Proc. ICASSP '88*, New York, NY (1988), pp. 48-51.

15. D.B. Paul, "The Lincoln Continuous Speech Recognition System: Recent Development and Results," in *Proc. DARPA Speech and Natural Language Workshop*, Philadelphia, PA (1989), pp. 160-166.

16. D.B. Paul, "The Lincoln Robust Continuous Speech Recognizer," in *Proc. ICASSP '89*, Glasgow, Scotland (1989), pp. 449-452.

17. D.B. Paul, "Tied Mixtures in the Lincoln Robust CSR," in *Proc. DARPA Speech and Natural Language Workshop*, Cape Cod, MA: Morgan Kaufmann Publisher (1989), pp. 293-302.

18. D.B. Paul "A CSR-NL Interface Specification," in *Proc. DARPA Speech and Natural Language Workshop*, Cape Cod, MA: Morgan Kaufmann Publisher (1989), pp. 203-214.

19. R.C. Rose and D.B. Paul, "A Hidden Markov Model Based Keyword Recognition System," to be published in proc. ICASSP '90.

20. R.P. Lippmann, "Review of Neural Networks for Speech Recognition," *J. of Neural Computation 1*, 1-38, (1989).

21. R.P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, Vol. 4, 4-22 (1987).

22. R.P. Lippmann, "Neural Nets for Computing," in *Proc. ICASSP '88*, New York, NY (1988), pp. 1-6.

23. R. P. Lippmann, "Neural Network Classifiers for Speech Recognition," *The Lincoln Laboratory Journal*, Vol. 1, No. 1, 107-124 (1988).

24. Y. Lee and R.P. Lippmann, "Practical Characteristics of Neural Network and Conventional Pattern Classifiers on Artificial and Speech Problems," in *Proc. IEEE Neural Information Processing Systems (NIPS) Conference*, Boulder, CO, to be published.

25. W.Y. Huang and R.P. Lippmann, "HMM Speech Recognition with Neural Net Discrimination," in *Proc. IEEE NIPS*, Boulder, CO, to be published.

26. F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," in *Proc. IEEE*, Vol. 73, No. 11, 1616-1624 (1985).

27. S.E. Levinson, "Structural Methods in Automatic Speech Recognition, " in *Proc. IEEE*, Vol. 73, No. 11, 1625-1650 (1985).

28. L.R. Rabiner and B.H. Juang, "An Introduction to Hidden Markov Models," *IEEE Acoust. Speech Signal Process. Magazine*, Vol. 3, No.1, 4-16 (1986).

29. J.K. Baker, "The Dragon System - An Overview," *IEEE Trans. Acoust. Speech Signal Process. Magazine* , Vol. ASSP- 23, No. 1, pp. 24-29 (1975).

30. G.R. Doddington and T.B. Schalk, "Speech Recognition: Turning Theory into Practice," *IEEE Spectrum*, 26-32 (1981).

31. S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously-Spoken Sentence," *IEEE Trans. Acoust. Speech Signal Process. Magazine*, Vol. ASSP-28, No. 4, 357-366 (1980).

32. Proc. '86 DARPA Speech Recognition Workshop, (1986); Science Applications International Corporation Report SAIC-86/1546 (1986).

33. Proc. '87 DARPA Speech Recognition Workshop (1987); Science Applications International Corporation Report SAIC-87/1644 (1987).

34. Proc. '89 DARPA Speech and Natural Language Workshop, Cape Cod, MA: Morgan Kaufmann Publisher (1989).

35. Proc. '89 DARPA Speech and Natural Language Workshop, Cape Cod, MA: Morgan Kaufmann Publisher (1989).

36. R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context-Dependent Modelling for Acoustic-Phonetic Recognition of Continuous Speech," in Proc. *ICASSP '85*, Tampa, FL (1985), pp. 1205-1208.

37. K.F. Lee and H. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM," in *Proc. ICASSP '88*, New York, NY (1988), pp. 123-126.

38. K.F. Lee, *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers, 1989 (contains tutorial material and a thorough bibliography on HMM).

39. H. Murveit and M. Weintraub, "1000-word Speaker-Independent Continuous Speech Recognition Using Hidden Markov Models," in *Proc. ICASSP '88*, New York, NY (1988), pp. 115-118.

40. F. Kubala, Y. Chow, A Derr, M. Feng, O. Kimball, J. Makhoul, P. Price, R. Rohlicek, S. Roucos, R. Schwartz, and J. Vandergrift, "Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000-word Resource Management Database," in *Proc. ICASSP '88*, New York, NY (1988), pp. 291-294.

41. P. Price, W. Fischer, J. Bernstein, and D. Pallett, "The DARPA 1000-word Resource Management Database for Continuous Speech Recognition," in *Proc. ICASSP '88*, New York, NY (1988), pp. 651-654.

42. P.K. Rajasekaran, G.R. Doddington, and J.W. Picone, "Recognition of Speech Under Stress and in Noise," in *Proc. ICASSP '86*, Tokyo, Japan (1986), pp. 733-736.

43. L. Netsch, A. Smith, G. Doddington, and P. Rajasekaran, "Robust Recognition of Speech under Stress and Noise," in *Proc. National Electronic Convention (NAECON)* (1988).

44. E. Lombard, "Le Signe de l'Elevation de la Voix," *Ann. Maladiere Oreille*, Larynx, Nez, Pharynx, Vol. 37 1911, pp. 101-119.

45. D. Pisoni, R.H. Bernacki, H.C. Nussbaum, and M. Yuchtman, "Some Acoustic-Phonetic Correlates of Speech Produced in Noise," in *Proc. ICASSP '85*, Tampa, FL (1985), pp. 1581-1584.

46. R.K. Moore, M.J. Russell, and M.J. Tomlinson, "The Discrimination Network: a Mechanism for Focusing Recognition in Whole-Word Pattern Matching," in *Proc. ICASSP '83*, Boston, MA (1983), pp. 1041-1044.

47. B.J. Stanton, L.H. Jamieson, and G.D. Allen, "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions," in *Proc. ICASSP '88*, New York, NY (1988), pp. 331-334.

48. H. Bourlard, N. Morgan, and C.J. Wellekens, "Statistical Inference in Multilayer Perceptions and Hidden Markov Models with Applications in Continuous Speech Recognition," Technical Report, Philips Research Laboratory, Brussels, 1989.

49. M.A. Franzini, M.J. Witbrock, and K. F. Lee, "A Connectionist Approach to Continuous Speech Recognition," in *Proc. ICASSP '89*, Glasgow, Scotland (1989), pp. 425-428.

50. M. Russell and R. Moore, "Explicit Modelling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition," in *Proc. ICASSP '85*, Tampa, FL (1985), pp. 5-8.

51. J.M. Baker and D.F. Pinto, "Optimal and Suboptimal Training Strategies for Automatic Speech Recognition in Noise and the Effects of Adaptation on Performance," in *Proc. ICASSP '86*, Tokyo, Japan, (1986), pp. 745-748.

52. J.R. Bellagarda and D.H. Nahamoo, "Tied Mixture Continuous Parameter Models for Large Vocabulary Isolated Speech Recognition," in *Proc. ICASSP '89*, Glasgow, Scotland (1989), pp. 13-16.

53. X.D. Huang and M.A. Jack, "Semi-Continuous Hidden Markov Models for Speech Recognition," *Computer Speech and Language*, (1989).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	26 February 1990	

4. TITLE AND SUBTITLE

Robust Speech Recognition Using Hidden Markov Models: Overview of a Research Program

5. FUNDING NUMBERS

C — F19628-90-C-0002
PE — 62301E
PR — 337

6. AUTHOR(S)

C.J. Weinstein, D.B. Paul, and R.P. Lippmann

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Lincoln Laboratory, MIT
P.O. Box 73
Lexington, MA 02173-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

Technical Report 875

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

DARPA/ISTO
1400 Wilson Blvd.
Arlington, VA 22209

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ESD-TR-89-280

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report presents an overview of a program of speech recognition research which was initiated in 1985 with the major goal of developing techniques for robust high-performance speech recognition under the stress and noise conditions typical of a military aircraft cockpit. The work on recognition in stress and noise during 1985 and 1986 produced a robust Hidden Markov (HMM) isolated word recognition system with 99 percent speaker-dependent accuracy for several difficult stress/noise databases, and very high performance for normal speech. Robustness techniques which were developed and applied include: multi-style training; robust estimation of parameter variances; perceptually-motivated stress tolerant distance measures; use of time-differential speech parameters; and discriminant analysis. These techniques and others produced more than an order-of magnitude reduction in isolated word recognition error rate relative to a baseline HMM system. An important feature of the Lincoln HMM system has been the use of continuous-observation HMM techniques, which provide a good basis for the development of the robustness techniques, and avoid the need for a vector quantizer at the input to the HMM system. Beginning in 1987, the robust HMM system has been extended to continuous speech recognition for both speaker-dependent and speaker-independent tasks. The robust HMM continuous speech recognizer was integrated in real-time with a stressing simulated flight task, which was judged to be very realistic by a number of military pilots. Phrase recognition accuracy on the limited-task-domain (28-word vocabulary) flight task is better than 99.9 percent. Recently, the robust HMM system has been extended to large-vocabulary continuous speech recognition, and has yielded excellent performance in both speaker-dependent and speaker-independent recognition on the DARPA 1000-word vocabulary resource management database. Current efforts include further improvements to the HMM system, techniques for the integration of speech recognition with natural language processing, and research on integration of neural network techniques with HMM.

14. SUBJECT TERMS			15. NUMBER OF PAGES
robust speech recognition	discriminant analysis	neural networks	46
HMM speech recognition	speech recognition	speaker-dependent-recognition	16. PRICE CODE
hidden Markov model	simulated stress	speaker-independent-recognition	
	continuous speech recognition	multi-style training	

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	