NASA Contractor Report 177547 USAAVSCOM Technical Memorandum 90-A-001

Evaluation of Speech Recognizers for Use in Advanced Combat Helicopter Crew Station Research and Development

Carol A. Simpson

(NASA-CR-177547) EVALUATION OF SPEECH RECOGNIZERS FOR USE IN ADVANCED COMBAT HELICOPTER CREW STATION RESEARCH AND DEVELOPMENT (Psycho-Linguistic Research Associates) 90 p N90-24265
CSCL 01C G3/05 Unclas 0289158

CONTRACT NAS2-12425
March 1990

National Aeronautics and Space Administration

US ARMY AVIATION SYSTEMS COMMAND
AVIATION RESEARCH AND TECHNOLOGY ACTIVITY
MOFFETT FIELD, CA 94035-1099

NASA Contractor Report 177547 USAAVSCOM Technical Memorandum 90-A-001

Evaluation of Speech Recognizers for Use in Advanced Combat Helicopter Crew Station Research and Development

Carol A. Simpson

Psycho-Linguistic
Research Associates
Woodside, California

Prepared for
Ames Research Center
CONTRACT NAS2-12425
March 1990

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035-1000

US ARMY
AVIATION
SYSTEMS COMMAND

AVIATION RESEARCH AND
TECHNOLOGY ACTIVITY
MOFFETT FIELD, CA 94035-1099

CONTENTS

EVALUATION OF SPEECH RECOGNIZERS FOR USE IN ADVANCED COMBAT HELICOPTER CREW STATION RESEARCH AND DEVELOPMENT

Carol A. Simpson
Psycho-Linguistic Research Associates
Woodside, California

SUMMARY

The U.S. Army Crew Station Research and Development Facility uses vintage 1984 speech recognizers. An evaluation was performed of newer off-the-shelf speech recognition devices to determine whether newer technology performance and capabilities are substantially better than that of the Army's current speech recognizers. The Phonetic Discrimination (PD-100) Test was used to compare recognizer performance in two ambient noise conditions: quiet office and helicopter noise. Test tokens were spoken by males and females and in isolated-word and connected-word mode. Better overall recognition accuracy was obtained from the newer recognizers. The report lists recognizer capabilities needed to support the development of human factors design requirements for speech command systems in advanced combat helicopters.

INTRODUCTION

Psycho-Linguistic Research Associates (PLRA) is under contract to the U.S. Army at Ames Research Center, Moffett Field, California, to provide human factors support for speech command design for crew station research and development for future Army combat helicopters. Under the auspices of this contract PLRA designed a versatile simulation test-bed, called the Smart Command Recognizer (SCR) for speech command research. This test-bed is part of the U.S. Army Crew Station Research and Development Facility (CSRDF) at Moffett Field and is described in refs 10 and 11. It uses speech recognition devices that were available in 1984.

In May 1988, PLRA was directed by the Army's Crew Station Research and Development Branch to recommend whether the Army should upgrade the recognition devices used in the CSRDF, the supporting development and pilot training laboratories, and the AH-1S Cobra-based Flying Laboratory for Integrated Test and Evaluation (FLITE). In preparation for making this recommendation, PLRA has evaluated recent advances in off-the-shelf speech recognition technology to determine whether there is a substantial improvement in recognizer performance in the devices available today compared to the device that is currently used in the CSRDF and supporting laboratories. We estimate that between six and twelve devices would be required for this upgrade.

SYMBOLS

NC number of correct responses to legal words, i.e. words that are in the active vocabulary.

RA percent of legal words that are correctly recognized.

NJ number of correct rejections (no response) to illegal words, i.e. words that are not in the active vocabulary.

JA percent of illegal words that are correctly rejected.

NT total number of tokens presented for recognition; also the sum of the legal tokens (NL) and the illegal tokens (NI).

OA Overall Accuracy for both legal and illegal words.

$$OA = \frac{NC + NJ}{NT} \times 100$$

IR Insertion Rate of recognition responses when no token was presented.

$$IR = \frac{N\,[\text{insert}]}{NT}$$

AOA Adjusted Overall Accuracy, or Overall Accuracy adjusted for Insertions.

AF Adjustment Factor for Insertions.

$$AF = \left(1 - \frac{N\,[\text{insert}]}{NT} \right)$$

$$AF = 1 - IR$$

$$AOA = (OA)(AF)$$

$$AOA = \left(\frac{NC + NJ}{NT} \right) \left(1 - \frac{N\,[\text{insert}]}{NT} \right)$$

OBJECTIVE

The objective of this evaluation was to obtain sufficient information about performance and functional capabilities of current off-the-shelf recognition devices to serve as the basis for a recommendation to the U.S. Army's Crew Station Research and Development Branch (CSRDB) regarding upgrade of recognition technology. A finding of substantially better recognition performance, together with added functional capability for human factors research and development, will constitute grounds for a recommendation that the CSRDB initiate acquisition of upgraded speech recognition devices.

APPROACH

The evaluation was conducted in three phases. Phase I consisted of a survey of available off-the shelf systems to determine possible algorithm and device suitability for the CSRDF R&D mission. Phase II consisted of a Questionnaire Study of recognition device specifications and characteristics. Phase III was a rigorous performance evaluation of recognizers that passed the Phase II Questionnaire.

PHASE I

In Phase I, product literature was obtained from candidate vendors at trade shows, technical conferences, and via vendor advertising mailings. Those vendors whose literature appeared to offer speaker-dependent or speaker-independent connected word recognition in high ambient noise levels were selected for Phase II. As a result of Phase I, eleven vendors were selected to participate in Phase II; these vendors were AT&T Bell Laboratories, Crouzet, Dragon Systems, ITT Defense Communications, Kurzweil AI, Marconi, Smiths Industries*, Speech Systems Inc., Texas Instruments, Voice Control Systems, and Votan.

PHASE II

Phase II consisted of a Questionnaire Study. Participating vendors completed a Recognition Device Specifications and Characteristics Questionnaire, copy attached as Appendix A. The Questionnaire elicited information about each device relative to CSRDB recognition performance requirements, device-host interface requirements, and device function requirements. These questionnaires were analyzed to determine the vendors whose devices ranked the highest in meeting or exceeding CSRDB's minimum requirements. Five vendors, Crouzet (represented by Allied Signal), ITT, Marconi, Smiths Industries, and Votan, were invited

----

* Note that Lear-Siegler was acquired by Smiths Industries prior to this evaluation. Therefore the invitation was issued to Smiths only.

to participate in Phase III. Table I gives a summary of the capabilities of these five recognizers, according to the vendors' responses to the Phase II Questionnaire. For the Smiths recognizer, dimensions of the SIR-L rather than the SIR-T model are given since the SIR-L was the one supplied for the evaluation. The capabilities of the CSRDF recognizer, are given for comparison. Since the text describing the various capabilities has been highly abbreviated, the reader is urged to consult Appendix A for the full description of each capability. Potentially competition sensitive information is excluded from Table I.

TABLE I. - RESULTS OF PHASE II QUESTIONNAIRE

RECOGNIZER	CSRDF	CROUZET	ITT	MARCONI	SMITHS	VOTAN
MODEL	V-6050	Crouzet	VRS-1280	Macrospeak	SIR-L	VPC-2100
DIMENSIONS						
width	15.5"	< 18"	PC-Board	12.2"	22"	PC-Board
height	3.5"	< 6"		3.7"	11"	
depth	15"	< 18"		15.1"	30"	
WEIGHT	12 lb	< 30 lb	< 1 lb	16 lb	49 lb	< 1 lb
POWER	115V AC	28V DC	115V AC	115V AC	115 V AC	DC buss
HOST INTERFACE						
RS-232 serial, 9600 baud, software or hardware handshake	Yes	Yes	No	Yes	Yes	No
IBM PC buss with MS-DOS OS device drivers	No	No	No	No	No	Yes
RECOGNITION MODE						
speaker-dependent	Yes	Yes	Yes	Yes	Yes	Yes
speaker-independent	No	No	No	No	Yes	No*
connected word	Yes	Yes	Yes	Yes	Yes	Yes
connected word across nodes	Yes	Yes	Yes	Yes	Yes	Yes
TEMPLATE ENROLLMENT						
max 2 tokens	Yes	No	No	Yes	Yes (DTW)	Yes
can delete individual templates	Yes	Yes	Yes	Yes	Yes	Yes
can re-enroll individual words	Yes	Yes	Yes	Yes	Yes	Yes

----

* Votan offers a fixed vocabulary of 5 words in speaker-independent mode. While technically a speaker-independent capability, this is too limited to handle the 200- to 1000-word vocabularies used in the CSRDF.

TABLE I CONTINUED

	CSRDF	CROUZET	ITT	MARCONI	SMITHS	VOTAN
goodness of fit data for 1st & 2nd best match during template testing	Yes	Yes	Yes	Yes	No	Yes
other type of comparative data on templates during testing	No	No	Yes	Yes	Yes	No
audio input level feedback during enrollment	Yes	Yes	No	No	No	Yes
non-volatile, pilot-portable low-cost template storage	Yes	No	No	Yes	?*	No
AUDIO INPUT CONTROLS AND DISPLAYS						
input gain	Yes	Yes	Yes	Yes	Auto	Yes
choice of mic or line level	Yes	Yes	Yes	Yes	Yes	Yes
display input level during recognition	No	TBD	No	No	Yes	No
audio monitor jack for signal actually received by recognizer	No	Yes	No	No	No	No
TOTAL VOCABULARY SIZE (approximate)						
total vocabulary	255	400	500	640	800	300
active vocabulary	50	?	> 100	> 100	96	300
VOCABULARY STRUCTURE						
min 50 nodes	Yes	Yes	1500	400	256	Yes
any word can be in 1 to all nodes concurrently	Yes	Yes	Yes	Yes	Yes	Yes
can change vocabulary independently of node structure	Yes	Yes	Yes	Yes	Yes	Yes
can change node structure independently of vocabulary	Yes	Yes	Yes	Yes	Yes	Yes
run-time redefinition of nodes	No	No	No	No	No	Yes

----

* Smiths offers a pilot-portable, non-volatile storage data module, but cost and availability are unknown at time of writing.

TABLE I CONTINUED

	CSRDF	CROUZET	ITT	MARCONI	SMITHS	VOTAN
RECOGNITION ALGORITHM CONTROLS						
real-time acceptance threshold control	Yes	N/A	Yes	Yes	Yes	Yes
programmable time-out	Yes	?	No	Yes	Yes	Yes
time-out can be defeated under program control	Yes	?	Yes	Yes	Yes	Yes
recognition mode abort	Yes	Yes	Yes	Yes	Yes	Yes
RECOGNITION ALGORITHM OUTPUT						
word number recognized	Yes	Yes	Yes	Yes	Yes	Yes
template number recognized	No	TBD	Yes	N/A	N/A	Yes
peak audio input level of token just recognized connected	No	No	No	No	No	No
goodness of fit for 1st & 2nd closest match during connected word recognition	No	No	No	Y/N (1st only)	No	Yes
WORD RECOGNITION ACCURACY (Vendors' own reports, PLRA report for CSRDF)						
min 80% for connected digits spoken in 80 to 90 dB SPL of noise	No	?	Yes	Yes	Yes	Yes
min 95% for words spoken as two-word phrases in 80 to 90 dB SPL of noise	No	Yes	Yes	Yes	Yes	Yes

TABLE I CONTINUED

	CSRDF	CROUZET	ITT	MARCONI	SMITHS	VOTAN
RECOGNITION RESPONSE TIME (Vendors' own reports, PLRA report for CSRDF)						
max 500 ms from end of spoken word to recognizer response	No	Yes	Yes	Yes	Yes	Yes
max of 100 ms from end of spoken word to recognizer response	No	Yes	Yes	?*	Yes	Yes
max of 1 sec. from end of 3rd word of 3-word phrase to recognizer response	Yes	Yes	Yes	?*	Yes	Yes
max of 200 ms from end of 3rd word of 3-word phrase to recognizer response	No	?	Yes	?*	Yes	No
AVAILABILITY						
Off-the-shelf as of 15 August 1988	No (out of prod)	Yes	No (MS-DOS)	Yes	Yes	Yes
Capable of 30 days delivery of up to 12 units as of 1 December 1988	No	No	No	?	No	Yes

----

* Marconi preferred not to answer these questions without more explicit definition of the acoustic definition of end of word.

Terms of Phase III Participation

Vendor participation in Phase III was subject to certain terms and conditions.

Each candidate vendor agreed to bail to PLRA one device for purposes of PD-100 testing. Any required maintenance was provided by the vendor. Each vendor was invited to send one technical staff member to ensure that, within the structure of the PD-100 Test procedure, the enrollment of templates was done according to the vendor's recommendations.

Each vendor agreed that all test data will remain the sole property of PLRA and PLRA will retain the sole right to decide whether to release and/or to publish the test data, in whole or in part. PLRA agreed that, except for reporting to the Army the names of the devices, if any, which performed substantially better than the current CSRDF device, any release or publication by PLRA of the actual test data would be de-identified with respect to vendor or device name.

PLRA agreed that an individual vendor could, at any time during the PD-100 testing, elect to withdraw from the evaluation if in that vendor's opinion, provided in writing to PLRA, the test was not being conducted fairly with equal treatment of all recognition devices.

PLRA agreed to provide each participating vendor with an advance copy of PLRA's report to the Army containing PLRA's recommendations to the Army regarding the possible upgrade of the Army's Crew Station Research and Development Facility Speech I/O Testbed and including the names of those devices, if any, that exhibited substantially better performance than the current CSRDF device. Each vendor was invited to comment, in writing, on the report, and PLRA agreed to reproduce and include the comments of each vendor, unedited, in PLRA's final report to the Army. Each vendor's comments, together with any proprietary information regarding that vendor's device, will be provided in a separately bound proprietary Appendix for that vendor, and will not be available for public distribution. Each participating vendor will receive a copy of the final report, including that vendor's own comments.

Phase III Performance Evaluation

All five vendors agreed to the terms and conditions of the Phase III evaluation.

Phase III began August 29, 1988, and consisted of a rigorous recognition performance assessment of the candidate devices selected during Phase II. The final results of Phase III together with the demonstrated device functions constitute the basis for

PLRA's recommendation to the CSRDB regarding upgrade.

The recognition performance measurement was performed at PLRA's speech testing lab using the Phonetic Discrimination (PD-100) Test, developed by PLRA, for rigorous assessment of speech recognition accuracy (See Refs 1-3).

PD-100 Test Measures

The traditional approach to evaluation of speech recognizers was impacted greatly when the PD-100 Test was introduced in 1987. With its introduction, a phonetic discrimination assessment method was added to the existing methods for recognition evaluation. Two concepts are crucial to the PD-100 Test. The first concept involves the assessment of the recognizer's phonetic discrimination ability for minimal phonetic differences in words. The second concept involves a systematic and well-defined procedure for measuring a recognizer's resistance to false alarms.

Phonetic Discrimination

The speech recognition research and development (R&D) community has long known that the effect of acoustic similarity among active vocabulary items is stronger than sheer active vocabulary size on recognition accuracy for legal words (NRC, 1984). Armstrong in 1980 while investigating the effects of performing a manual pursuit tracking task on recognition accuracy, found the detrimental effect on recognition accuracy due to phonetic similarity of the test tokens to be much stronger than the effect of the added motor task (ref 8). The Alvey Project in the UK and the European Economic Community (EEC) ESPRIT Project, both concerned with the establishment of standards, assessment procedures, databases, and common tools for multilingual speech response systems, have advocated the use of phonetically balanced test materials. The UK Alvey Project (Taylor, 1988, ref 12) developed a Speech Technology Assessment (STA) technique which addresses, among other parameters, the effects of phonetic range on speech recognition accuracy. In the most comprehensive report on the subject, known to this author, the ESPRIT Project 1541 Final Report (ref 9) reviews efforts on a world-wide basis and recommends the development of a multi-lingual data base and a set of methodologies and tools for the consistent use of this data base across different recognizers and different languages (Barry, Harland, Hazan, and Fourcin, 1988).

The PD-100 Test addresses phonetic discrimination by controlling for the phonetic distance among test words. Phonetic distance is measured in terms of the number of minimal phonetic features that are different among pairs of words in the test set. The design and derivation of the PD-100 Test are given in refs 1-3. It includes a sub-set of the word pairs in the Diagnostic Rhyme Test (DRT) by Voiers, et al (ref 7). To these are added other words which themselves contain phonemes (speech sounds) that are not included in the DRT. The result is a set of 100 words which contains all phonemes of English with all legal single consonants appearing in word initial and word final position and with all vowels, including diphthongs, represented. A sample of consonant clusters is also included. The PD-100 test words thus provide a more complete sample of the range English phonetic segments than does the DRT. The hundred words are divided into two half-lists such that each word in a given list has a mate in the other list which differs from it by a particular number of phonetic features. These word pairs are further assigned to sub-lists, based on the phonetic distance between them.

For testing, templates for one member of each pair, in a given sub-list, are put into the recognizer's active vocabulary. These words become the legal words for the test. The other member of each pair of words is put into the list of illegal words, as illustrated in Figure 1.

ACTIVE VOCABULARY	TEST VOCABULARY	
VEAL	VEAL	LEGAL
	FEEL	ILLEGAL
BEAN	BEAN	LEGAL
	PEAN	ILLEGAL
•	•	
•	•	
•	•	
MEAT	MEAT	LEGAL
	BEAT	ILLEGAL
NIP	NIP	LEGAL
	DIP	ILLEGAL

Figure 1. - Assignment of word pairs to legal and illegal lists

The best overall recognition accuracy is obtained for those recognizers that exhibit the highest rate of correct recognition of the legal words and also the highest rate of correct rejection of the illegal words.

Advantage of PD-100 Test Difficulty

Because some of the PD-100 sub-lists contain word pairs that differ by only one phonetic feature, the test is very difficult, even for human listeners. Other sub-lists for which the phonetic distance between pairs is greater are less difficult but still not expected to permit perfect overall recognition accuracy. The extreme difficulty of the PD-100 test ensures that no recognizers will even approach perfect performance. A group of recognizers might all score extremely well on an easier test, but this would preclude any valid statistical tests of differences due to the skewed distribution and a ceiling effect. This same group of recognizers may score in the mid range of possible PD-100 scores making the discovery of differences among them more likely because of the greater sensitivity of statistical tests when applied to data that are drawn from a normal distribution unconstrained by ceiling effects.

Rejection Accuracy

Traditionally, recognizers have been tested for recognition accuracy, by presenting for recognition only words that are in the active vocabulary set. The active vocabulary set is " ...the (instantaneously varying) subset of [the words or phrases to be recognized] that may be active at a given time because of an imposed task grammar or other syntactic constraint,..." (Pallett, NBS, 1984 (ref 4)). The PD-100 Test, in addition to measuring "Recognition Accuracy" as the ratio or percent of "legal" words recognized, also measures "Rejection Accuracy" as the ratio or percent of "illegal" words correctly rejected. Legal words are words that are in the recognizer's active vocabulary/ies, and illegal words are words that are not in the active vocabulary/ies. "Overall Accuracy" is a function of both the number of legal tokens of words that are properly recognized and of the number of tokens of illegal words that are properly rejected.

By presenting both legal and illegal tokens, one can measure overall accuracy. Further, by manipulating the degree of phonetic similarity between the legal and illegal tokens, one can assess phonetic discrimination down to the level of individual phonetic features. Figure 2 illustrates the test token types and response types that are covered by the PD-100 Test procedure. The multi-level arrow, labeled "Phonetic Similarity", indicates that overall accuracy will vary as a function of phonetic similarity of the legal and illegal test tokens.

Prior to the introduction of the PD-100 Test, rejection accuracy was seldom addressed in recognition performance assessment (see Williamson, and Curry (ref 6) for one of the few assessments of rejection accuracy) and never addressed at the single-feature phonetic discrimination level provided by the PD-100 Test. Good rejection accuracy is just as important as good recognition accuracy. A recognizer should respond correctly to those words which are in its active vocabulary and should not produce false alarms to words which are acoustically similar but not in its active vocabulary. System designers cannot depend on users, even cooperative users, to speak only those words which are in the active vocabulary.

The PD-100 Test generates a measure called "Overall

Accuracy". Overall Accuracy includes both recognition accuracy and rejection accuracy. Both must be good in order for Overall Accuracy to be good. The formula for Overall Accuracy is given in Simpson and Ruth, 1987 (ref 1) as

$$OA = \frac{NC + NJ}{NT} \times 100$$

OA is Overall Accuracy.
NC is the number of correct responses to legal words, i.e. words that are in the active vocabulary.
NJ is the number of correct rejections (no response) to illegal words, i.e. words that are not in the active vocabulary.
NT is the total number of tokens presented for recognition and is also the sum of the legal tokens (NL) and the illegal tokens (NI).

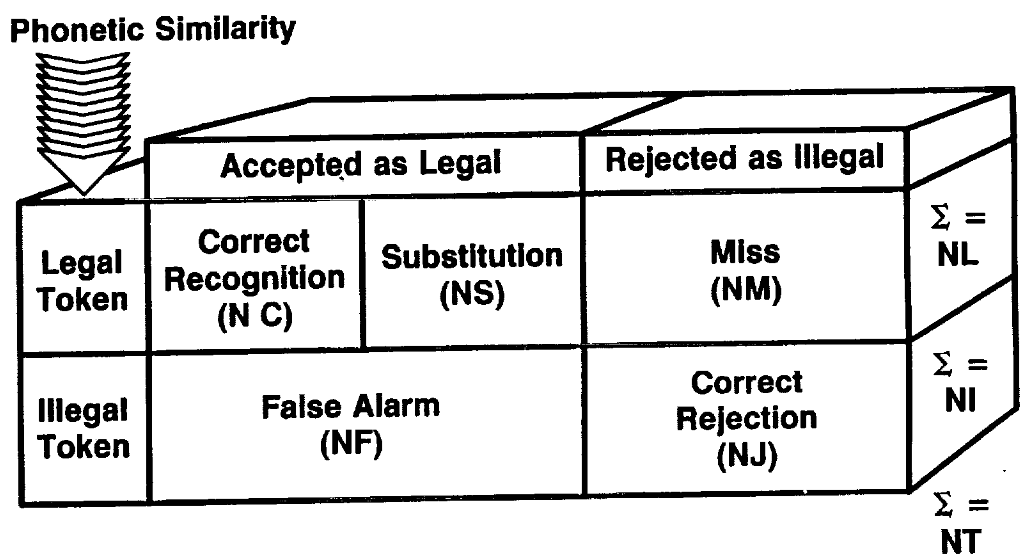

Figure 2. - PD-100 Speech Recognition Response Matrix

Overall Accuracy, as defined in 1987, does not handle what are traditionally called "insertion errors", i.e. recognition responses to non-speech sounds such as coughs, environmental noise, etc. (Pallett, 1984 (ref 4)). Since insertions, like false alarms, are frustrating to the user and slow down the process of accomplishing tasks via speech command, they are of concern to our evaluation for the CSRDB. Pilots cannot tolerate extra time to correct recognition errors due to either insertions or false alarms.

During previous PD-100 Test applications we have recorded spurious recognition responses that were not associated with the presentation of any of the legal or illegal test word tokens in the PD-100 data base. Therefore, the PD-100 Test was recently upgraded to evaluate and account for errors in speech recognition due to insertions. In consultation with Dr. John C. Ruth, consulting mathematician to PLRA through DEV AIR Technical Associates, we have developed an adjustment to Overall Accuracy to account for insertions.

The Adjustment Factor has the following characteristics.

1. If the number of insertions, N[insert] is zero, the original Overall Accuracy is retained.

2. If the number of insertions, N[insert], is equal to the number of test tokens (NT), the original overall accuracy is reduced to zero.

3. If the number of insertions, N[insert], is between zero and NT, the overall accuracy is diminished in the same proportion as the Insertion Rate (IR), defined as:

$$IR = \frac{N\,[\text{insert}]}{NT}$$

4. If the number of insertions, N[insert], is greater than NT, a negative value is generated for the Adjusted Overall Accuracy (AOA). The larger the negative value, the poorer is the performance of the recognizer in rejecting insertions.

The Adjustment Factor (AF) is represented by the function:

$$AF = \left(1 - \frac{N\,[\text{insert}]}{NT} \right)$$

or

$$AF = 1 - IR$$

The Adjustment Factor (AF) is applied to the Overall Accuracy in the following manner. Adjusted Overall Accuracy = (Overall Accuracy) x (Adjustment Factor) or

$$AOA = (OA)(AF)$$

or

$$AOA = \left(\frac{NC + NJ}{NT} \right) \left(1 - \frac{N\,[\text{insert}]}{NT} \right)$$

In this manner the effect of insertions can be represented by the degradation of the overall accuracy or, in extreme cases, as a negative value for AOA. A negative value for AOA will alert a developer or experimenter that the total number of insertions compared to total test tokens is unacceptable.

Phase III Experimental Design

This evaluation used four speakers of the available sixteen speakers in the PD-100 speech token database - two males and two females with a General American dialect.* The test token variables included 1) manner of speaking: single words versus connected words in phrases, 2) speech distortion: normal voice versus muffled voice versus extreme pitch and rate variations; 3) sex of speaker: male versus female; and 4) acoustic environment: quiet versus helicopter noise. For each combination of test token variables, the four PD-100 measures described above were computed from the data: Adjusted Overall Accuracy (AOA), Overall Accuracy (OA), Recognition Accuracy (RA), and Rejection Accuracy (JA).

AOA for isolated PD-100 words was measured for each of the four speakers for each of two noise conditions: quiet laboratory, and tape recorded UH-60 helicopter noise, the same noise that is modeled in the CSRDF. Measured sound pressure level in the laboratory was 57-58 dB SPL (ref. 0.0002 dyne/cm2) with all recognizers and associated host computers operational. The UH-60 cockpit noise was presented at 85 dB SPL. All speech tokens were presented at 100 dB SPL, plus or minus 5 dB intra-speaker variability among tokens. Therefore the signal-to-noise ratio for the quiet condition averaged +42 to +43 dB compared to +15 dB for the UH-60 noise condition.

In addition to the isolated PD-100 word tests, connected PD-100 word phrases were used to test connected word recognition and word spotting capabilities. The connected word tests were conducted for all four speakers, in the quiet condition only.

For one of the males and one of the females, data were also collected in quiet for muffled speech tokens and for speech tokens spoken with extreme values of pitch (high and low) and of rate (fast and slow).

All the recognizers received exactly the same speech token at the same time via a custom audio distribution system which permitted adjustment of signal level independently for each device to provide the manufacturer's recommended input signal level for that device. The simultaneous presentation of each speech token to all recognizers eliminated variability for a given token

----

* Other speakers in the database represent Eastern and Southern American English dialects, British English, and German accented English.

between different presentations.

A human listener listened to and responded to each test token or test phrase at the same time that it was presented, in parallel, to the recognizers. The human provided a benchmark against which the recognizers could be compared.

Phase III Procedure

Set-up and enrollment of the Phase III devices was scheduled between August 29 and November 11, 1988. Each vendor was scheduled individually for one of six (6) week-long time periods, according to the original schedule shown below.

29 AUG - 2 SEPT 1988	ENROLLMENT - MARCONI
19 SEP - 23 SEP 1988	ENROLLMENT - SPARE
26 SEP - 30 SEP 1988	ENROLLMENT - VOTAN
24 OCT - 28 OCT 1988	ENROLLMENT - SMITHS
31 OCT - 4 NOV 1988	ENROLLMENT - CROUZET/ ALLIED BENDIX
7 NOV - 11 NOV 1988	ENROLLMENT - ITT

Four of the five vendors actually participated in the evaluation. Allied Signal, representing Crouzet, notified PLRA three days before they were scheduled to begin enrollment that they had very reluctantly, for financial reasons, decided to put their speech recognition program on hold. They stressed that this decision in no way reflected upon their respect for the Crouzet device. PLRA, with concurrence from Allied Signal, then invited Crouzet to participate directly, but Crouzet reluctantly declined to participate alone without its U.S. partner.

Of the four remaining vendors, all but Marconi required more than one week to complete the set-up, enrollment, and template verification and calibration. The latter half of November and the first part of December were used to complete the preparation for Votan, ITT, and Smiths; and finally on December 14, 1988, enrollment for all recognizers was completed.

Our original estimate of technical staff support from each vendor was one week with a maximum of two weeks. Actual staff support needed to configure the recognizers for vocabulary enrollment and the PD-100 test syntax varied from a low of 2 days to a high of 3 weeks, depending on the user interface, development tools, and device functions of the recognition device. The enrollment process to enroll and verify templates for 100 words for each of four speakers in the PD-100 speech token database required as little as 6 hours and as much as 6 days for the different devices.

Details of the enrollment process for the CSRDF recognizer are given in Appendix B. Details of the enrollment process for the ITT, Marconi, Smiths, and Votan recognizers are given in Appendices C, D, E, and F, respectively. After reviewing their respective appendices, each vendor had the option to keep all or any part of the appendix in the proprietary section of this final report. All vendors decided to permit their respective appendices to be published in full in the non-proprietary section of the final report.

Some vendors chose also to have their responses to the final report published in the non-proprietary section of the report. The non-proprietary responses are included as Appendix G.

Those responses that are considered proprietary by individual vendors are separately bound and may be distributed only to those government individuals who have a need to know their contents in conjunction with a potential government procurement.

Template Enrollment

The template enrollment procedure was designed to ensure that the templates for each recognizer were created according to the recommended practices of the vendors, within the constraints of the PD-100 Test. The following constraints were imposed.

Constraints on Template Enrollment. - A maximum of two tokens per PD-100 word was allowed for practical reasons. The research schedule at CSRDF with operational Army pilots and visiting research pilots does not allow for lengthy enrollment sessions. Previous experience with systems using more than two tokens has taught us that pilots become fatigued and frustrated with a resulting degradation in recognition performance.

Templates were enrolled in relative quiet, i.e. laboratory ambient noise. Again, our experience has been that pilots and experimenters alike become fatigued when exposed to simulated cockpit noise during the enrollment session. Therefore, we wanted to assess recognition performance for templates enrolled in quiet and then tested in quiet and in noise. We realize we could obtain better accuracy for templates enrolled in noise but do not want to pay the price for this in terms of fatigue. Additionally, we want one set of templates to be usable in different noise backgrounds. We have had relatively good success with templates enrolled in quiet using the current CSRDF device, provided we test the templates and re-enroll any poor ones prior to use and provided we take care to have a clean audio distribution system with repeatable signal levels and good signal-to-noise ratios. We have also found it is critical to train the pilots how to talk to the recognizer, just as they have to learn how to fly a particular helicopter. Both are, for them, relatively overlearned motor control tasks. With instruction and sufficient real-time feedback pilots do adapt to new aircraft. Similarly, we have found that instruction and real-time feedback on speech performance enables pilots to adapt their speech production to some extent.

PD-100 Template Enrollment Procedure. - The procedure for enrolling templates for the PD-100 Test includes several steps, listed below:

- Enroll Practice Vocabulary
    - Select Gain for Best Performance
    - Select Acceptance Threshold
        - For Quiet environment
        - For Noise environment
- Test PD-100 Active Set Control (Syntax)
- Test Data Collection

- Enroll PD-100 Vocabulary for each speaker
    - Initial Enrollment
    - Test Templates
    - Re-enroll as desired
    - Calibrate Templates

A practice vocabulary of 100 words, consisting of the numbers from one to one hundred, e.g. one, two, three ... ten, eleven, twelve ..., twenty, twenty-one, ... ninety-eight, ninety-nine, one hundred, is used for set-up purposes. This vocabulary has the same number of words as the PD-100, and has words which are phonetically similar. It provides an opportunity to program the test device with a vocabulary file of size 100 and to program the syntax nodes that will be needed during PD-100 testing to control the active vocabulary selections. It is also used for determining the best input signal level, device input gain, acceptance threshold, etc. without using the actual test vocabulary. Use of the practice vocabulary removes the danger of customizing the device settings to the actual test vocabulary and provides for a routine enrollment of the test vocabulary.

For each of the five recognizers, then, all set-up and checkout of the device was done using the practice vocabulary. The user's manuals in combination with recommendations from the vendor's technical representatives provided guidance as to the best structure for the syntax, within the constraints of the PD-100 Test.

Each device was programmed to enroll a single vocabulary of 100 words. Set definition commands were used for each device to divide the 100 word vocabulary into the three lists, List 1, List 2, and List 3 of the PD-100 Test. In this way, a single master set of templates for the 100 words could be made for each device, providing further experimental control of the test and greatly reducing the template enrollment time, as compared to the alternative of enrolling the lists separately. This procedure also exercised the capabilities of each device for syntax or set definition, switching of the active set, and independent manipulation of syntax and vocabulary with respect to templates.

Additionally, set definition commands were used to create half-lists of each of the three PD-100 Lists. The half-lists, called List 1a, List 2a, and List 3a, contained exactly half of the words of each parent list. These words were the words that would be made legal, i.e. be the active recognition set, during PD-100 Testing.

Templates were verified by testing them with the tokens that had been used to create them and using an extra set of tokens that were not used for enrollment but were also not the test tokens. At no time prior to the actual data collection were the test tokens presented to the recognizers, and the vendors' technical representatives never heard the test tokens. This precaution was taken to ensure that we would not inadvertently tune the templates for the particular test tokens we used.

We allowed the technical representative for each vendor to re-enroll templates for words which were not correctly recognized during the template verification. They were allowed to use reasonably simple techniques that would be likely to be employed by a moderately experienced user of speech recognition devices but which did not require detailed technical knowledge of the recognition algorithm. These techniques included using a different enrollment token (four were available to chose among), adjusting the recognition input gain using available user controls, and positioning the token delivery audio tape. The different vendors took advantage of re-enrollment to varying degrees. Smiths witnessed the verification for one of the four speakers, made no changes, and left the verification and calibration of the other three speakers' templates to PLRA. Marconi re-enrolled two words for one speaker and otherwise made no changes. Votan analyzed recognition performance on the practice vocabulary of the numbers from one to one hundred. On the basis of the practice numbers vocabulary accuracy, Votan elected to leave all template enrollment to PLRA. ITT actively participated in the entire enrollment process and re-enrolled from 6 to 35 words for each of the four speakers. Whatever the level of involvement by the vendor's technical representative, each one eventually reached a point at which he or she announced that continued re-enrollment would not likely result in better accuracy.

Each vendor was asked to select a rejection threshold to be used during the test. Since overall accuracy is a function of both recognition accuracy and rejection accuracy, we wanted to be sure to have an appropriate rejection threshold that would, in the vendor's judgment, maximize adjusted overall accuracy. Vendors were permitted to select one rejection threshold for use in quiet and, if desired, a different one for use during the tests in UH-60 noise. As an aid to selecting the threshold for noise, they observed recognition performance on the numbers vocabulary in signal-to-noise ratios of +15 dB and +10 dB of simulated piston engine, two-bladed helicopter noise.

Each vendor's normal procedure for recognition in noise was used, with the restriction noted above that templates had to be enrolled in quiet. Details of these methods are described in the respective appendices for the individual vendors.

Smiths, Votan, and Marconi, as well as the CSRDF recognizer, met the criterion of a maximum of two tokens per PD-100 word. ITT normally requires more than two tokens per word. In a test performed at PLRA, they attempted to reduce their enrollment procedure to meet the 2-token limit. In the opinion of the ITT technical representative, performance was considerably worse than it would have been had the full procedure been used, requiring four or more tokens per word, with some of those tokens embedded in phrases. Therefore, a compromise was reached, so as to include ITT in the evaluation. However, ITT understood that the CSRDF recognizers must be capable of good performance with only two tokens. The full enrollment procedure was used for two of the PD-100 speakers: M1 and F2 (one male and one female). The templates for the other two speakers, male M3 and female F4, were bootstrapped from those of M1 and F2, respectively, and were made using only two isolated tokens per word.

Data Recording

All recognizer responses were recorded in two separate computer files for redundancy and cross-check. One of these files was generated automatically and consisted of an ASCII text transcript of each response for a particular recognizer, generated by the recognizer user interface software and written to the data file. This software had to be written explicitly for this evaluation using available application program development software tools supplied by the vendor. In the case of the CSRDF device, the Smart Command Recognizer (SCR) software performed the data file recording function.

The CSRDF, Votan VPC-2100, and ITT VRS 1280 systems were programmed to create their own data files. The CSRDF and the Votan were able to produce MS-DOS format data files directly since they were operating under the MS-DOS operating system. The ITT system provided for evaluation operated under the xenix operating system because the MS-DOS version of ITT's software was not yet fully operational. Therefore a conversion program, supplied by ITT, was used to convert the xenix format files to MS-DOS format. Again, ITT understood that the use of the xenix version was only a means to testing their algorithm and that the CSRDF requirement remains for a system that can function within the MS-DOS environment. The

Marconi Macrospeak was programmed to display its responses to its terminal. These responses were then captured by a terminal emulation program running on a Marconi-supplied IBM-PC compatible computer and were written into an MS-DOS format data file by the terminal emulation program. The Smiths system was intended to operate in the same manner as the Marconi. However, we were unsuccessful in our attempts to interface its terminal display serial port to a second IBM-PC compatible computer. Therefore, the redundant data file for the Smiths consisted of a hand-written data log.*

The composite data for all recognizers and the human listener were collected in a separate text file on a separate computer. This file was created and edited during data collection by one of the Experimenters and was formatted as required by the data analysis program which computes Overall Accuracy and its associated measures.

Two Experimenters observed the five recognizers and listened to the human listener as they all responded to the test tokens. The test tokens were presented one word or phrase at a time. Then data was recorded in the composite data file for each recognizer, including the human listener. The Experimenters cross-checked each other's reports. The human listener, who was stationed in a remote location, reported what he heard via intercom and used each word in a short phrase to ensure that his responses were correctly perceived by the Experimenters. After the data collection, the composite data were verified against the individual data files for each of the recognizers.

Phase III Data Analysis and Reporting

Phase III data collection was completed on December 18, 1988. This final report, sent to vendors for review in July 1989, covers data analysis for the normally spoken isolated words in quiet and in noise and for connected word phrases in quiet. Vendors also received copies of the PD-100 data, summarized by speaker and condition, for their respective recognizers. These data will be discussed with individual vendors to provide diagnostics and observations on the strengths and weaknesses of their own recognizers relative to the CSRDF requirements.

Phase III Results

Results are presented separately for single words and for

----

* Subsequently, during data analysis, a second attempt was made to interface the Smith's serial port to an IBM-PC compatible computer. This was successful, and trial recognition data from the Smiths recognizer were successfully captured and stored in an MS-DOS format data file.

connected words.

Single Word Results

Data for the single words, spoken in normal voice by the four speakers, and presented in quiet and in noise, were analyzed in two ways. First the data were analyzed using each vendor's selected acceptance threshold. Then the data were analyzed, for those recognizers that were able, using the optimized acceptance threshold. The optimized threshold was computed for OA with the aid of the STRAP program (also called SRET), developed at Wright-Patterson Air Force Base (Williamson, 1988 (ref 5)). The Marconi and the Votan were the only recognizers that provided goodness of fit scores during connected recognition. Therefore, the optimization could only be done for these two devices.*

Analysis Procedure. - For both the vendor's choice and the optimized threshold, the analysis procedure was identical. The human performance scores were used as a benchmark to normalize the scores of the five recognizers. This was done because the human's AOA and OA were substantially better than that of the recognizers. We wanted to test for significant differences among recognizers, not between the human and the recognizers. The human's average AOA, across speakers, was 93.0 +/-2.94 in quiet and 93.5 +/-3.87 in noise. Since the human had no insertion errors, AOA and OA were identical. The average AOA and average OA of all five recognizers, across speakers, after normalization via the human scores, is shown below:

	VENDOR'S CHOICE		OPTIMIZED THRESHOLD	
	Quiet	Noise	Quiet	Noise
AOA	61.5	57.8	66.0	59.5
OA	63.9	59.6	68.3	61.3

Analysis of Variance (AOV) was used to test for significant differences among recognizers. A Three-Way AOV was performed on the entire single word data set with three variables of Recognizer, Environment, and Sex of speaker. There were five recognizers (CSRDF, ITT, Marconi, Smiths, Votan) by 2 levels of environment (quiet, noise) by 2 levels of sex (male, female).

----

*In discussions with ITT after their review of the final report, it was determined that an alternative method may have been available to obtain goodness of fit scores from their recognizer. Had optimized thresholds been determined for the ITT recognizer, the results might have been better or worse than those reported here. For details, refer to Note 1, following the References section.

Speakers were treated as subjects in the analysis. The Recognizer and Environment were within subjects variables, while Sex was a between subjects variable.

This AOV was performed for each of the four PD-100 Test measures: Adjusted Overall Accuracy (AOA), Overall Accuracy (OA), Recognition Accuracy (RA), and Rejection Accuracy (JA).

Due to the relatively small sample size of speakers, the confidence level of 0.10 was used to determine statistical significance.

Single Words, Vendor's Choice Threshold. - For AOA, the effect of recognizer in the 3-way AOV was not significant. However, the interaction of recognizer and environment was significant ($F=3.52$, $df=4,8$; $p < 0.10$). No other variables or interactions were significant.

For OA, the effect of recognizer was significant ($F=3.09$, $df=4,8$; $p < 0.10$). Environment (quiet versus noise) was also significant ($F=12.92$, $df=1,2$; $p < 0.10$), and the interaction between environment and recognizer was significant ($F=5.37$, $df=4,8$; $p < .05$). No other effects or interactions were significant for OA.

For RA, the effect of environment was highly significant ($F=38.52$, $df=1,2$; $p = 0.025$) as was the interaction of recognizer and environment ($F=7.05$, $df=4,8$; $p < 0.025$). No other effects or interactions were significant.

For JA, the effect of recognizer was significant ($F=4.39$, $df=4,8$, $p < 0.05$) as were the effect of environment ($F=17.91$, $df=1,2$; $p < 0.10$) and especially the interaction between recognizer and environment ($F=59.70$, $df=4,8$; $p \ll 0.001$).

Because of the strong interaction between recognizer and environment, the two halves of the data base (quiet and noise) were next tested separately. And, in order to determine which recognizers might have performed significantly better than the CSRDF recognizer, Duncan's Multiple Range Test was then used. First, a One-Way AOV was performed on the AOA data for the quiet condition. The effect of differences among recognizers was highly significant ($F=4.51$, $df=4,12$; $p < 0.025$). Then, the results of the Duncan's Multiple Range Test for the quiet condition indicated that the ITT, the Marconi, and the Smiths recognizers performed better than the CSRDF recognizer at the 0.05 confidence level. The Votan did not perform significantly better in quiet for the Vendor's choice threshold than did the CSRDF recognizer ($p > 0.10$).

A similar analysis of differences between the four evaluation recognizers compared to the CSRDF recognizer for the data collected in noise showed no significant differences among any of the recognizers for Vendor's choice threshold in noise.

Thus, for isolated PD-100 words presented in quiet, there were three recognizers with Adjusted Overall Accuracy significantly better than the CSRDF recognizer. But, for those same words by the same speakers presented in UH-60 cockpit noise with a signal-to-noise ratio of +15 dB, there were no significant differences among recognizers in Adjusted Overall Accuracy, for Vendor's Choice acceptance threshold.

Single Words, Optimized Threshold. - The 3-way AOV was performed for the five recognizers after the data for the Marconi and the Votan had been changed to reflect the results obtained with the optimized threshold. The scores for both these recognizers improved with the use of the optimized threshold.

For AOA, the effect of recognizer in the 3-way AOV was significant ($F=4.06$, $df=4,8$; $p < .05$), as was the interaction of recognizer and environment ($F=3.77$, $df=4,8$; $p < 0.10$). No other variables or interactions were significant.

For OA, the effect of recognizer was significant ($F=3.97$, $df=4,8$; $p < 0.05$). Environment (quiet versus noise) was also significant ($F=10.40$, $df=1,2$; $p < 0.10$), and the interaction between environment and recognizer was significant ($F=6.54$, $df=4,8$; $p < .05$). No other effects or interactions were significant for OA.

For RA, the effect of environment was highly significant ($F=57.46$, $df=1,2$; $p = 0.025$) as was the interaction of recognizer and environment ($F=7.63$, $df=4,8$; $p < 0.01$). No other effects or interactions were significant.

For JA, the effect of recognizer was significant ($F=3.07$, $df=4,8$, $p < 0.10$), and the interaction between recognizer and environment was highly significant ($F=41.38$, $df=4,8$; $p << 0.01$).

As with the Vendor's Choice Threshold data, there was a strong interaction between recognizer and environment. So, the two halves of the data base (quiet and noise) were next tested separately. And, in order to determine which recognizers might have performed significantly better than the CSRDF recognizer, Duncan's Multiple Range Test was then used. First, a One-Way AOV was performed on the AOA data for the quiet condition.

The effect of differences among recognizers was highly significant ($F=6.01$, $df=4,12$; $p < 0.01$). Then, the results of the Duncan's Multiple Range Test for the quiet condition indicated that the ITT, the Marconi, the Votan, and the Smiths recognizers all performed better than the CSRDF recognizer at the 0.05 confidence level, using the optimized threshold.

A similar analysis of differences between the four evaluation recognizers compared to the CSRDF recognizer for the data collected in noise showed a significant difference among the recognizers for optimized threshold in noise ($F=2.81$; $df=4,12$; $p < 0.10$). Comparisons among means using the Duncan's Multiple Range Test showed one recognizer, ITT, to have performed worse in noise than the CSRDF at the .05 confidence level, with no significant differences between each of the other three recognizers and the CSRDF recognizer.

In summary, under quiet test conditions, three recognizers achieved significantly better Adjusted Overall Accuracy than the CSRDF recognizer. When the optimized threshold was used for two of the recognizers, the AOA for these recognizers improved, with the result that all four recognizers performed significantly better than the CSRDF recognizer. When the same test words by the same speakers were presented in UH-60 cockpit noise with a signal-to-noise ratio of +15 dB, one recognizer performed significantly worse than the CSRDF with no other significant differences between each of the other three recognizers and the CSRDF recognizer in Adjusted Overall Accuracy. This failure to perform better in noise than the CSRDF recognizer was exhibited for both the vendor's choice threshold and for the optimized acceptance threshold.

Table II shows the results of the comparisons for the Single Words. "Better" indicates significantly better performance than the CSRDF recognizer at $p < 0.10$. "*" indicates that an optimized threshold was obtained for that recognizer.

TABLE II. - COMPARISON OF FOUR RECOGNIZERS TO THE CSRDF RECOGNIZER AOA QUIET AND NOISE FOR VENDOR'S THRESHOLD AND OPTIMIZED THRESHOLD

	QUIET		NOISE	
	VENDOR'S THRESHOLD	OPTIMIZED THRESHOLD	VENDOR'S THRESHOLD	OPTIMIZED THRESHOLD
ITT VRS-1280	Better	Better		
Marconi Macrospeak*	Better	Better		
Smiths SIRL	Better	Better		
Votan VPC-2100*		Better		

Connected Word Results

Data for the connected words, spoken in normal voice by the four speakers, and presented in quiet were also analyzed using the Vendor's Choice threshold and using the Optimized threshold. The optimized threshold was computed to maximize OA, as described in the section above on Single Word Results. As with the single words, the Marconi and the Votan were the only recognizers that provided goodness of fit scores during connected recognition.* Therefore, the optimization could only be done for these two devices.

Analysis Procedure. - For both the vendor's choice and the optimized threshold, the analysis procedure was identical. The single word performance in quiet was included in this analysis for comparison to connected word performance. The human performance scores were used as a benchmark to normalize the scores of the five recognizers. The human's average AOA, across speakers, was 99.2 +/-0.92 for connected words and 93.0 +/-2.94 for single words. Since the human had no insertion errors, AOA and OA were identical. The average AOA and average OA of all five recognizers, across speakers, after normalization via the human scores, is shown below:

	VENDOR'S CHOICE		OPTIMIZED THRESHOLD	
	Connected	Single	Connected	Single
AOA	50.6	61.5	53.3	66.0
OA	51.6	63.9	54.2	68.3

Analysis of Variance (AOV) was used to test for significant differences among recognizers. A Three-Way AOV was performed on the entire quiet data set with three variables of Recognizer, Speaking Mode, and Sex of speaker. There were five recognizers (CSRDF, ITT, Marconi, Smiths, Votan) by 2 levels of speaking mode (connected, single word), by 2 levels of sex (male,female). Speakers were treated as subjects in the analysis. The Recognizer and Speaking Mode were within subjects variables, while Sex was a between subjects variable.

This AOV was performed for each of the four PD-100 Test measures: Adjusted Overall Accuracy (AOA), Overall Accuracy (OA), Recognition Accuracy (RA), and Rejection Accuracy (JA).

As with the single word analysis, the confidence level of 0.10 was used to determine statistical significance.

----

* However, see the footnote regarding the ITT recognizer in the section on Single Word Results.

Connected Words, Vendor's Choice Threshold. - For AOA, the effect of recognizer in the 3-way AOV was highly significant ($F=7.21$, $df=4,8$; $p < .01$), as was the effect of connected versus single word speaking mode ($F=30.92$, $df=1,2$; $p < 0.05$). No other variables or interactions were significant.

For OA, the effect of recognizer was highly significant ($F=8.23$, $df=4,8$; $p < 0.01$). Speaking Mode was also significant ($F=39.79$, $df=1,2$; $p < 0.05$), and the interaction between speaking mode and recognizer was significant ($F=6.14$, $df=4,8$; $p < .05$). No other effects or interactions were significant for OA.

For RA, the effect of speaking mode was significant ($F=13.44$, $df=1,2$; $p < 0.10$) as was the effect of recognizer ($F=6.09$, $df=4,8$; $p < 0.05$). No other effects or interactions were significant.

For JA, the effect of recognizer was highly significant ($F=29.75$, $df=4,8$, $p << 0.01$) as was the effect of the interaction between recognizer and speaking mode ($F=4.09$, $df=4,8$; $p < 0.05$).

The two halves of the data base (connected word and single word) were split and the connected word data tested separately using a 1-way AOV. The single word data in quiet had already been tested, as described in the section above on Single Word Results. Duncan's Multiple Range Test was then used in order to determine which recognizers might have performed significantly better than the CSRDF recognizer.

First, a One-Way AOV was performed on the AOA data for the connected condition. The effect of differences among recognizers was highly significant ($F=8.18$, $df=4,12$; $p < 0.01$). Then, the results of the Duncan's Multiple Range Test for the quiet condition indicated that the ITT, the Marconi, the Votan, and the Smiths recognizers all performed better than the CSRDF recognizer at the 0.05 confidence level for connected words in quiet using the vendor's choice threshold.

Thus, for connected PD-100 words presented in quiet there were four recognizers with Adjusted Overall Accuracy significantly better than the CSRDF recognizer, using Vendor's Choice acceptance threshold.

Connected Words, Optimized Threshold. - The 3-way AOV was next performed for the five recognizers after the data for the Marconi and the Votan had been changed to reflect the results obtained with the optimized threshold. The scores for both these recognizers improved with the use of the optimized threshold.

For AOA, the effect of recognizer in the 3-way AOV was significant ($F=12.16$, $df=4,8$; $p < .01$), as was the effect of speaking mode ($F=11.88$, $df=1,2$; $p < 0.05$). No other variables or interactions were significant.

For OA, the effect of recognizer was significant ($F=12.05$, $df=4,8$; $p < 0.01$). Speaking mode (connected versus single word) was also significant ($F=15.60$, $df=1,2$; $p < 0.10$), No other effects or interactions were significant for OA.

For RA, the effect of recognizer was significant ($F=7.22$, $df=4,8$; $p < 0.01$) as was the effect of speaking mode ($F=14.49$, $df=4,8$; $p < 0.10$). No other effects or interactions were significant.

For JA, the effect of recognizer was highly significant ($F=26.47$, $df=4,8$, $p \ll 0.10$), but, interestingly, not the effect of speaking mode ($F=0.08$, $df=1,2$; $p > 0.10$). No other effects or interactions were significant.

The two halves of the data base (connected and single word) were next split and the connected word data tested separately, using a 1-way AOV, followed by Duncan's Multiple Range Test.

The effect differences among recognizers was highly significant ($F=10.23$, $df=4,12$; $p < 0.01$). Then, the results of the Duncan's Multiple Range Test for the connected words in quiet indicated that the ITT, the Marconi, the Votan, and the Smiths recognizers performed better than the CSRDF recognizer at the 0.05 confidence level, using the optimized threshold data.

In summary, all four recognizers performed better than the CSRDF recognizer for the connected PD-100 words presented in quiet. These four exhibited Adjusted Overall Accuracy that was significantly better than the CSRDF recognizer. These results are shown in Table III. The results of single words in quiet are included for comparison. "Better" indicates significantly better performance than the CSRDF recognizer at $p < 0.10$. "*" indicates that an optimized threshold was obtained for that recognizer.

TABLE III. - COMPARISON OF 4 RECOGNIZERS TO CSRDF RECOGNIZER
AOA IN QUIET FOR SINGLE AND CONNECTED WORDS
USING VENDOR'S THRESHOLD AND OPTIMIZED THRESHOLD

	SINGLE WORDS		CONNECTED WORDS	
	VENDOR'S THRESHOLD	OPTIMIZED THRESHOLD	VENDOR'S THRESHOLD	OPTIMIZED THRESHOLD
ITT VRS-1280	Better	Better	Better	Better
Marconi Macrospeak*	Better	Better	Better	Better
Smiths SIRL	Better	Better	Better	Better
Votan VPC-2100*		Better	Better	Better

DISCUSSION

The results indicate that there are commercially available recognizers which provide significantly higher Adjusted Overall Accuracy than the CSRDF recognizer in quiet conditions for both single words and connected words. None of the four recognizers performed significantly better than the CSRDF recognizer in noise at +15 dB S/N. The one recognizer which performed worse than the CSRDF suffered from a higher insertion error rate during the noise trials than during the trials in quiet. This accounts in part for its lower Adjusted Overall Accuracy.

The reason for the interaction between the two experimental conditions of recognizer and environment (S/N) is not known. However, we can speculate that it has to do with the particular signal-to-noise ratios selected for this evaluation. In quiet conditions (average S/N of 42.5 dB), all four recognizers had higher AOA than the CSRDF recognizer. At S/N of +15 dB, none of the recognizers achieved a higher AOA than the CSRDF recognizer. This suggests there is a crossover signal-to-noise ratio at which some of the recognizers would have achieved higher Adjusted Overall Accuracy than the CSRDF recognizer. The actual crossover S/N might well be different for each recognizer. Just where these cross-over S/N's might be, however, cannot be determined from the data collected in this evaluation. The graph in Figure 3 illustrates the concept of cross-over S/N ratios that would be specific to individual recognizers. It does not portray any actual data.

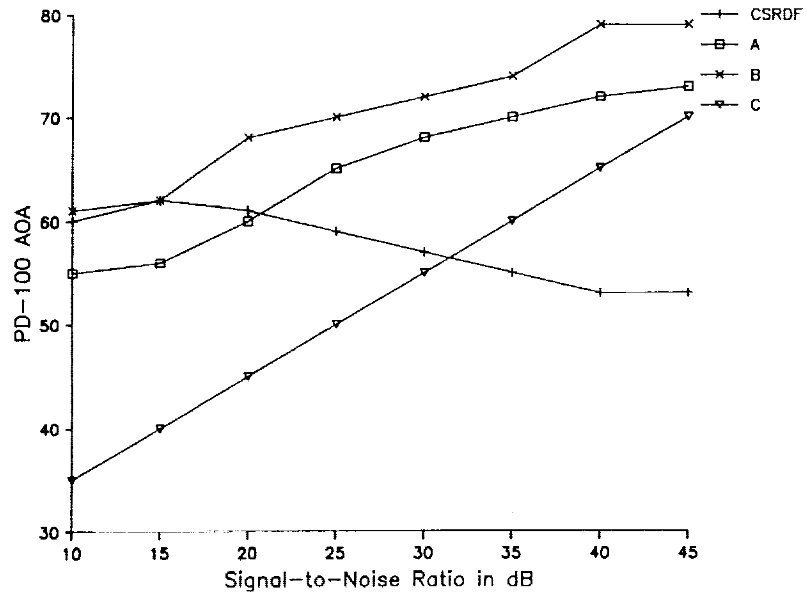

Figure 3. - Theoretical Effect of Signal-to-Noise ratio on Adjusted Overall Accuracy for PD-100 Test

The CSRDF is normally operated at cockpit noise levels no greater than 75 dB SPL. Assuming a normal 100 dB SPL speaking rate at the pilot's boom microphone, a S/N of +25 dB can be achieved. Depending on the value of the theoretical crossover S/N, a S/N of +25 dB may be sufficient to obtain the better performance analogous to that exhibited in this evaluation by the newer recognizers in quiet at S/N of +42 dB.

Table IV summarizes the results of human-normalized Adjusted Overall Accuracy by comparing the AOA for the CSRDF recognizer to the mean of those other recognizers which performed as well as or better than the CSRDF, using optimized threshold data for the Marconi and the Votan. The mean for performance in quiet is thus based on four recognizers. The mean for performance in noise is based on only three recognizers and does not include the AOA for the recognizer which performed worse than the CSRDF recognizer in noise.

TABLE IV.- HUMAN NORMALIZED AOA FOR CSRDF 1984 VINTAGE RECOGNIZER COMPARED TO THE MEAN FOR FOUR 1988 VINTAGE RECOGNIZERS

	CSRDF	MEAN	N of RECOGNIZERS
Single Words S/N = +42	53	69	4
Connected Words S/N = +42	40	57	4
Single Words S/N = +15	62	61	3

It is also worth noting in Table IV that AOA for the CSRDF recognizer was better in noise than in quiet. The converse was true for the mean of the other four recognizers. An analysis of the recognition accuracy for legal words and the rejection accuracy for illegal words, for the CSRDF, revealed that the CSRDF's better AOA in noise was mainly due to a reduction of false alarms in noise, compared to the false alarm rate in quiet. It will be remembered that the CSRDF acceptance threshold was set fairly wide open for that device at 50 in quiet and was stopped down to 27 in noise. Indeed, in quiet conditions, the CSRDF suffered from a high false alarm rate. Human-normalized rejection accuracy (the converse of false alarms) was a mere 10 in quiet compared to 69 in noise. The combined effect of the noise and the tighter acceptance threshold was to reduce false alarms while also reducing, but to a lesser degree, recognition accuracy for legal words.

Acceptance threshold optimization clearly has value in that Adjusted Overall Accuracy was improved for the two recognizers which provided distance or goodness of fit scores. It is useful to note that the vendors' chosen thresholds were not optimal. We also found that the value of the optimum threshold was different for different speakers and for different noise conditions. Recognition was also more difficult for one of the four speakers, Speaker M3, than for the other three, perhaps because of more variability in the amplitude and speaking rate of his speech. The ability to quickly determine the optimum threshold for individual pilots would be of high value for CSRDF operations.

It was hoped that the current recognition technology would have demonstrated better performance in noise than the CSRDF recognizer, which is now four years old. As suggested above in the discussion on cross-over S/N, perhaps a signal-to-noise ratio of +15 dB was too difficult for all the recognizers when the requirement was also for very good phonetic discrimination; a signal-to-noise ratio of +20, +25, or +30, more typical of industrial and office environments where speech recognition has made considerable progress, might have demonstrated an advantage of today's technology over the older CSRDF recognizer. In spite of the disappointing results in noise, there are other reasons which support a CSRDF upgrade. All four newer recognizers performed better in quiet conditions, for both single words and connected words, in a very difficult phonetic discrimination test than did the CSRDF recognizer. Particularly those two recognizers, the Votan VPC-2100 and the Marconi Macrospeak, that output the goodness of fit or distance score during connected recognition offer the possibility of easily optimizing the acceptance threshold for individual pilots and for modifying the acceptance threshold during recognition as a function of goodness of fit data and in response to pilots' speech variability.

All four recognizers offer larger total vocabularies, larger active vocabularies, and faster recognition response times than the CSRDF recognizer. Three of the four, Smiths, Votan, and Marconi are compatible with the CSRDF MS-DOS operating system environment. Two of these, Marconi and Votan, provide the necessary flexibility of set structure to take advantage of the versatility and power of the CSRDB Smart Command Recognizer software.

One of the four, the Smiths, offers three different recognition algorithms in a single system together with exceptional research and development tools for the study of intra- and inter-speaker variability and its effects on speech recognition accuracy.

Given these advantages of the newer recognizers over the current CSRDF recognizer, and in consideration of the mission of the CSRDB to support the development of human factors design requirements for speech command systems for advanced combat helicopters, it is the author's recommendation that an upgrade to the CSRDB recognizers be undertaken. A minimum set of specifications is recommended in the section below on Minimum Specifications for CSRDF Recognizer.

RECOMMENDED MINIMUM SPECIFICATIONS FOR CSRDF RECOGNIZERS

Must be capable of speaker-dependent or speaker-independent continuous connected word recognition across vocabulary nodes.

Must have a minimum of 500 word total vocabulary and an active vocabulary of at least 100 words.

Must provide at least 50 vocabulary nodes.

Documented phonetic discrimination performance in quiet equal to or better than the average performance exhibited by those recognizers in this evaluation, which performed better in quiet than the CSRDF recognizer and documented phonetic discrimination performance in noise equal to or better than the CSRDF recognizer*, as measured by the PD-100 test, specifically as shown in the table below:

	absolute	human-normalized
AOA for single words in quiet: (N = 4 recognizers)	65	69
AOA for connected words in quiet: (N = 4 recognizers)	56	57
AOA for single words in noise: at a S/N of +15 dB or worse for helicopter noise (= 3 recognizers)	58	61

Minimum PD-100 AOA Scores Required for CSRDF Recognizer

Performance within one standard deviation of the mean for the above conditions should be considered to meet the requirements. The corresponding standard deviations for the above means are given below:

	absolute	human-normalized
AOA for single words in quiet: (N = 4 recognizers)	3.3	3.5
AOA for connected words in quiet: (N = 4 recognizers)	4.3	4.4
AOA for single words in noise: at a S/N of +15 dB or worse for helicopter noise (N = 3 recognizers)	3.9	4.1

PD-100 Score Standard Deviations to be Used as Tolerances For CSRDF Recognizer Performance Specifications

Must demonstrate recognition response time for the third word of a three-word connected word phrase of no more than 200 ms from the end of the third word to the output of the recognized word for

----

*Recognizers which did not perform as well as the CSRDF recognizer for a particular condition are not included in the mean. The mean is used since statistical analysis was performed only to determine those recognizers which performed better or worse than the CSRDF recognizer and not to determine comparative performance among those recognizers.

active vocabulary set size of at least 50 words and with a set change occurring at least between the second and third words.

Compatibility with the CSRDB Smart Command Recognizer software and hardware in at least one of the following ways. 1) Recognizer controllable for all functions via RS-232 port, 9600 baud; 2) Recognizer controllable for all functions via IBM-PC buss and with vendor-supplied device drivers callable in 'C', compatible with the Microsoft C compiler, version 4.x or 5.x, and running under the MS-DOS operating system.

Compatibility with the CSRDB Smart Command Recognizer syntax interpreter in all of the following ways: 1) any given vocabulary word may reside concurrently in any 1 or more nodes; 2) active set change can be individually specified for each word in the active set such that each word can cause a jump to a different set; 3) active set change can be controlled directly by a host during ongoing connected word recognition; 4) vocabulary, assignment of words to sets, and templates can each be changed independently of any of the other two and without destroying the recognizer memory's copy of the other two.

Must meet the above phonetic discrimination performance specifications using no more than two isolated tokens per vocabulary word for enrollment.

Must permit deletion and re-enrollment of individual templates and of templates for individual words without affecting other already enrolled words and templates.

Must be capable of enrolling templates that meet the above performance specifications for a given individual for 100 words in no more than one hour total time for enrollment, including any time needed for re-enrollment of problem words.

Must provide a capability for sampling ambient noise and using this information during recognition to account for noise which was not present at time of template enrollment.

Must output goodness of fit score for each word recognized, as each word is recognized (not waiting until end of command), during connected, continuous recognition.

Must provide the following under program control: 1) input gain adjustment 2) acceptance threshold adjustment, 3) active set change forced by host program. Active set change must be accomplished in response to host directive fast enough to accommodate connected word mode of speaking during set change.

Must provide non-volatile, pilot-portable, low-cost template storage.

Must provide input gain under program control and choice of mic or line level input.

If a recognition time-out is incorporated, then this time-out must be programmable for duration and also be defeatable under program control.

Must be powered by either 115 V AC if a stand-alone unit or receive power from the IBM-PS buss if a board for IBM-PC compatible computers.

Vendor must be able to repair or replace malfunctioning units within 30 working days and must be able to supply loaner units during the repair period.

ACKNOWLEDGMENTS

The enthusiastic participation of the four recognizer vendors, ITT Defense Communications, Marconi Electronics, Smiths Industries, and Votan, contributed greatly to the success of this evaluation. Special thanks are extended to the technical representatives of each participant, who worked long hours at PLRA during the system set-up and template enrollment phases and contributed valuable expertise in the use of their respective products. They are Jeff Nichols, ITT; Les O'Leary, Marconi; Ian Bickerton, Smiths; and Sivia Loitz, Bruce Thompson, and Sanjeev Malaney, Votan. The SRET program, written by Systems Control Technology, Inc, Dayton OH, a copy of which was provided by David T. Williamson, AFWAL/FIGR, Wright-Patterson Air Force Base, was invaluable for the computation of optimum acceptance thresholds from the PD-100 data and saved literally hundreds of hours of labor. Jeffrey Chan, DEV AIR Technical Associates, serving as technical research assistant, designed and installed the custom audio distribution system, was my fellow experimenter, and performed much of the data reduction and analysis. This work was supported by NASA Ames Contract NAS2-12425 to Psycho-Linguistic Research Associates for support of advanced helicopter controls and displays for the U.S. Army Aeroflightdynamics Directorate at Ames Research Center, Moffett Field, CA. Dr. Nancy M. Bucher served as the Contract Technical Monitor.

REFERENCES

1. Simpson, C.A.; and Ruth, J.C.: The Phonetic Discrimination Test for Speech Recognizers: Part I. Speech Technology, April/May, 1987, pp. 47-53. New York: Media Dimensions. Also in Proc. of National Aerospace and Electronics Conf. NAECON 1987, Dayton OH, May 18-22, 1987, vol. 3, pp. 889-896. New York: IEEE.

2. Simpson, C.A.; and Ruth, J.C.: The Phonetic Discrimination Test for Speech Recognizers: Part II. Speech Technology, Oct/Nov,

1987, pp. 58-61. New York: Media Dimensions. Also in Proc. of National Aerospace and Electronics Conf. NAECON 1987, Dayton OH, May 18-22, 1987, vol. 3, pp. 889-896. New York: IEEE.

3. Simpson, C.A.; and Ruth, J.C.: Phonetic Discrimination Testing - A Fresh Approach to Applications Suitability Assessment, Proc. of Military Speech Tech '87, pp. 100-104. New York: Media Dimensions. Also in Proc. of 1987 Technical Meeting, Avionics Section, Air Armament Division, pp. 98-102. Washington, D.C.: American Defense Preparedness Association.

4. Pallett, D.S.: Performance Assessment of Automatic Speech Recognizers, J. of the National Bureau of Standards, vol. 90, no. 5, 1985, pp. 371-387.

5. Williamson, D.T.: Speech Recognition Evaluation Tool, Proc. of Military Speech Tech '88, pp. 102-106. New York: Media Dimensions.

6. Williamson, D.T.; and Curry, D.G.: Speech Recognition Performance Evaluation in Simulated Cockpit Noise, Proc. of Speech Tech '84, pp. 99-102. New York: Media Dimensions.

7. Voiers, W.D.; Cohen, M.E.; and Mickunas, J.: Evaluation of Speech Processing Devices, I. Intelligibility, Quality, Speaker Recognizability. Final Report. Contract No. AF19(628)4195, OAS (1965).

8. Armstrong, J.W.: The Effects of Concurrent Motor Tasking on Performance of a Voice Recognition System, Masters Thesis. Naval Postgraduate School, Monterey, CA, 1980.

9. Barry, B.; Harland, G.; Hazan, V.; and Fourcin, A.: Final Report, Definition Phase: 2.2.87-31.1.88, Multilingual Speech Input/Output Assessment, Methodology, and Standardization. ESPRIT Project 1541. University College London, 1988.

10. Simpson, C.A.; and Krones, R.R.: Smart Command Recognizer (SCR): A Rapid Prototyping System for Speech Commands. Proc. of Speech Tech '88, pp. 67-70. New York: Media Dimensions.

11. Simpson, C.A.; Bunnell, J.W.; and Krones, R.R.: Smart Command Recognizer (SCR): for Development, Test, and Implementation of Speech Commands. Proc. of AIAA Flight Simulation Technologies Conference, 1988, pp. 215-221.

Taylor, M.R.: Recognition Performance Prediction in Harsh Military Airborne Environments, Proc. of Military Speech Tech '88, pp. 107-111. New York: Media Dimensions.

NOTES

Note 1

In discussions with ITT during the preparation for testing, PLRA asked the ITT technical representative whether the numeric values output by the VRS-1280 along with the number and text for each recognized word were goodness of fit scores. The technical representative responded that ITT does not use goodness of fit scores to determine recognition acceptance but instead compares the value for the word recognized to the value obtained for a "rejection template" and decides to accept or reject depending on whether the word or the rejection template, respectively, received the lower value. This method, as understood by PLRA, does not use a fixed acceptance threshold. After reading the draft final report, ITT stated their belief that the values output during recognition, together with the values for the corresponding rejection templates, could have been treated as goodness of fit scores for purposes of analysis via STRAP. In discussions with ITT it was determined that ITT could have written their data recording program to accommodate the STRAP analysis requirements. Had such an analysis been performed, the results using the optimized threshold might have been better than, the same as, or worse than those obtained and reported here using the vendor-chosen method of comparing recognized word values to rejection template values.

Note 2

After reviewing the draft final report, ITT expressed concern that the syntax they had recommended for PD-100 testing had been designed to maximize recognition accuracy for isolated words with possible detrimental effects on connected word recognition. Specifically, in order to enhance word-boundary detection, their syntax caused the recognizer to look for a 400 ms pause in the signal before trying to recognize the next word. This would cause the recognizer to have a relatively higher miss rate for all but the initial words in connected word phrases. ITT stated they could have designed a syntax that would work for both isolated words and connected words. Had the test been conducted with a different syntax, the isolated word results might have been worse than, the same as, or better than those obtained and reported here.

APPENDIX A

Appendix A contains a copy of the questionnaire that was used for Phase II to gather detailed information from Phase II vendors regarding algorithm performance and device characteristics.

QUESTIONNAIRE

Recognition Device Specifications and Characteristics

INSTRUCTIONS

Complete this Questionnaire and return it to Psycho-Linguistic Research Associates at the address shown below, no later than June 20, 1988.

PSYCHO-LINGUISTIC RESEARCH ASSOCIATES
ATTN: Dr. Carol A. Simpson
485 Summit Springs Road
Woodside, California 94062 USA

There are two sections to this questionnaire. The first section covers minimum requirements needed and elicits information about device capabilities which exceed the minimum requirements. The second section covers highly desired capabilities.

Fill out the vendor and device information at the bottom of this page. Then complete Sections 1 and 2 of the questionnaire. If you have any questions about any item on the questionnaire, please call Dr. Carol Simpson at (415) 851-0917.

Vendor Name ____________________

Point of Contact ____________________

Address ____________________

____

____

Phone ____________________

Device Name/Model ______________

Qty. 1-12 price per unit to US Government (may be approximate) __________

QUESTIONNAIRE
Recognition Device Specifications and Characteristics

SECTION 1

MINIMUM CAPABILITIES

This section covers the minimum capabilities that are required for the recognition devices that are to be used in the research facility. Indicate for each item whether or not your device meets the minimum requirement by circling the YES or the NO. In addition, indicate the maximum capability of your device, in the column labeled MAXIMUM CAPABILITY. Then, on the blank lines below the individual items, list any additional capabilities that your device has for the general area covered by this set of items.

	MEETS MINIMUM (circle yes or no)		MAXIMUM CAPABILITY (provide details)
PHYSICAL CHARACTERISTICS			
Width: 18" or less (or 19" rack mount)	YES	NO	
Height: 6" or less	YES	NO	
Depth: 18" or less	YES	NO	
Weight: 30 lbs or less	YES	NO	
Power: 115-120 V	YES	NO	
Capable of reliable operation under military helicopter vibration and cockpit noise conditions; need not be certified flight-worthy.	YES	NO	

ADDITIONAL CAPABILITY: (PHYSICAL)

____

____

____

____

HOST INTERFACE	MEETS MINIMUM?		MAX CAPABILITY?
RS-232 serial, 9600 baud with hardware or software handshaking	YES	NO	
OR			
IBM-PC buss with MS-DOS operating system device drivers	YES	NO	

(Note: need only one of the above to meet minimum requirements)

ADDITIONAL CAPABILITY: (HOST INTERFACE)

____

____

____

____

RECOGNITION MODE	MEETS MINIMUM?		MAX CAPABILITY?
Speaker Dependent	YES	NO	
OR			
Speaker Independent	YES	NO	
Connected Word (no pauses required)	YES	NO	
Connected Word recognition across Vocabulary sets (nodes)	YES	NO	

ADDITIONAL CAPABILITY: (RECOGNITION MODE)

____

____

____

____

TEMPLATE ENROLLMENT	MEETS MINIMUM?	MAX CAPABILITY?
1 or at most 2 tokens required per vocabulary item	YES NO	
Deletion of individual templates	YES NO	
Re-enrollment of individual words without need to enroll other words	YES NO	
Distance or goodness of fit data for input token compared to 1st and 2nd closest match during template testing.	YES NO	
Audio level range detection and reporting during enrollment (level too high or too low)	YES NO	
Non-volatile, low cost, pilot-portable storage of templates, e.g. 3 1/4" disk	YES NO	

ADDITIONAL CAPABILITY: (TEMPLATE ENROLLMENT)

____

____

____

____

AUDIO INPUT CONTROLS	MEETS MINIMUM?	MAX CAPABILITY?
Input gain	YES NO	

RECOGNITION DEVICE QUESTIONNAIRE, PLRA 1988

ADDITIONAL CAPABILITY: (AUDIO INPUT)

____

____

____

____

TOTAL VOCABULARY SIZE	MEETS MINIMUM?	MAX CAPABILITY?
Minimum 250 words	YES NO	

ADDITIONAL CAPABILITY: (VOCAB SIZE)

____

____

____

____

VOCABULARY STRUCTURE	MEETS MINIMUM?	MAX CAPABILITY?
Minimum of 50 nodes	YES NO	
Any vocabulary item can reside in 1 to all nodes concurrently	YES NO	
Can change vocabulary independently of node structure	YES NO	
Can change node structure independently of vocabulary	YES NO	
Run-time modification of vocabulary structure (redefinition of nodes)	YES NO	

ADDITIONAL CAPABILITY: (VOCAB STRUCTURE)

____

____

____

____

RECOGNITION ALGORITHM CONTROLS	MEETS MINIMUM?		MAX CAPABILITY?
Real-time acceptance threshold control	YES	NO	
Programmable time-out for spoken input	YES	NO	
Time-out can be defeated under program control	YES	NO	
Recognition mode abort	YES	NO	

ADDITIONAL CAPABILITY: (ALGORITHM CONTROLS)

____

____

____

____

RECOGNITION ALGORITHM OUTPUT	MEETS MINIMUM?		MAX CAPABILITY?
Word number recognized	YES	NO	

ADDITIONAL CAPABILITY: (ALGORITHM OUTPUT)

____

____

____

____

RECOGNITION ACCURACY	MEETS MINIMUM?	MAX CAPABILITY?
min 80% word recognition accuracy for connected digits spoken by trained users in noise levels of 80 to 90 dB SPL. PLEASE SUPPLY SUPPORTING DATA.	YES NO	
min 95% word recognition accuracy for words spoken as two-word connected phrases by trained users in noise levels of 80 to 90 dB SPL. PLEASE SUPPLY SUPPORTING DATA.	YES NO	

ADDITIONAL CAPABILITY: (REC. ACCURACY)

____

____

____

____

RECOGNITION RESPONSE TIME	MEETS MINIMUM?	MAX CAPABILITY?
No longer than 500 ms to recognize one word, measured from end of word spoken by pilot to the return by the device of the word recognized, no longer than 500 ms.	YES NO	
No longer than 1 sec. to recognize third word of three-word connected phrases, measured from end of third word to return of third word recognized.	YES NO	

ADDITIONAL CAPABILITY: (REC. RESP. TIME)

____

____

____

____

AVAILABILITY	MEETS MINIMUM?	MAX CAPABILITY?
Off-the-shelf as of 15 AUGUST, 1988	YES NO	
Capable of 30 days delivery	YES NO	
of up to 12 units as of 1 DECEMBER 1988	YES NO	

ADDITIONAL CAPABILITY: (AVAILABILITY)

____

____

____

____

SECTION 2

HIGHLY DESIRED CAPABILITIES

This section of the Questionnaire covers capabilities which are not essential but are highly desired. Please indicate for each item whether your device provides the indicated capability. Please answer all items even if your responses in the section above already give the same information.

AUDIO INPUT CONTROLS AND DISPLAYS

Input resistance: choice of mic (ca. 250 ohms) or line (ca. 600-1000 ohms) (highly desired)	YES	NO
Display of audio input level during recognition, e.g. via VU-meter on the device front panel.	YES	NO
Audio monitor jack for headphone or line level output of audio signal that is received by the recognition device.	YES	NO

TOTAL VOCABULARY SIZE

500 words	YES	NO
1000 words	YES	NO

RECOGNIITON ALGORITHM OUTPUT

Template number recognized	YES	NO
Peak audio input level of token just recognized	YES	NO

Distance or goodness of fit data for input token compared to 1st and 2nd closest match, during connected word recognition. YES NO

RECOGNITION RESPONSE TIME

No longer than 100 ms to recognize one word, measured from end of word spoken by pilot to the return by the device of the word recognized, no longer than 500 ms. YES NO

No longer than 200 ms to recognize third word of three-word connected phrases, measured from end of third word to return of third word recognized. YES NO

APPENDIX B
PREPARATION OF THE CSRDF RECOGNIZER PD-100 TESTING

The CSRDF recognizer is a Votan Model 6050 stand-alone unit with two RS-232 serial ports, one for a terminal and another for two-way communications with a host computer. The unit includes a single 3 1/2 " floppy disk drive. It measures 3" x 15" x 15.5" and weighs 12 lb. Disk operating system software is supplied to load the recognition software from disk into the recognizer memory to read and write from and to disk, list the file directory, and copy files in single disk mode. The recognition software provides for operation in two modes: terminal mode via the terminal serial port and host peripheral mode via the host serial port. Hardware and software handshaking are available for the serial ports, selectable by dip switches inside the unit. Baud rates from 75 to 9600 are also selected by these internal dip switches. The terminal mode facilitates quick demonstrations of speech recognition. A structured, menu-driven program elicits vocabulary information and speech templates from the user. Several levels of a Help Menu provide interactive documentation. However, the host peripheral mode provides far more flexibility in order of template enrollment, changes to vocabulary and syntax, and control of the active recognition set during connected recognition. Therefore, the host peripheral mode is used for CSRDF operations.

User's Manuals

A user's manual called "VTR 6000 Users Guide" was supplied with the CSRDF recognizers when they were purchased by the Government in 1985. The manual lists in alphabetical order the commands for terminal mode operations and then, in alphabetical order those for the peripheral mode operations. It gives instructions for up and downloading templates via the host serial port. It also It lists the meaning of most of the error codes and includes an index.

Physical Installation

The CSRDF recognizer was interfaced to one of the CSRDB Smart Command Recognizer (SCR) computers, in the configuration in which it is normally used at the Crew Station Research and Development Branch laboratories and simulators.

Audio Interface

The CSRDF recognizer was connected to the PLRA audio distribution system line 3. The audio signal level was set to 1 mV RMS for a 1000 Hz tone at a presentation level of 94dB. The microphone level input of the CSRDF recognizer was used, as opposed to the line level input.

Software for Enrollment

Software for enrollment consisted of the CSRDB SCR "ENROLL" program. One vocabulary file containing the numbers vocabulary, one vocabulary file containing the PD-100 vocabulary, one enrollment sequence file containing the word numbers to be enrolled, and three set files - to divide the vocabulary into the three PD-100 Lists, were created with a text editor.

Software for Testing

Software for testing consisted of the CSRDB SCR "RECOG" program. Three more set files were created, one for each half-list of each of the three PD-100 Lists - Lists 1a, 2a, and 3a containing those words that would be in the active vocabulary during testing.

Noise Handling

The same two methods used for the Votan VPC-2100 noise handling were implemented for the CSRDF recognizer. The templates were enrolled at input gain level 3 and were tested in noise at input level 2. Additionally, the acceptance threshold was tightened from the normal level of 50 to 27 for the runs conducted in noise.

Software for Data Collection

Software for data collection was available as one of the functions of the SCR "RECOG" program. The program saves the word recognized, the active set number, and the time to the nearest second that the word was recognized. Additional data such as the second best match, the template number recognized are not provided by the CSRDF recognizer when it is in connected recognition mode, only when it is in isolated word recognition mode. All data were saved by the program to ASCII data files in MS-DOS format for later analysis.

Enrollment Procedure

<u>Numbers Vocabulary</u> - Initially, two templates were made for each word in the numbers vocabulary. However, when one of the three word lists was made active for recognition, e.g. List 1 with the numbers from 1 to forty, the template set vocabulary exceeded the recognizer's active vocabulary memory. Therefore, only one template per word was used for the checkout.

Templates were made for the numbers vocabulary at three input gains: level 2, level 3, and level 4 and were tested at these gains, respectively. The best recognition accuracy was obtained for templates made at gain 3 and tested at gain 3. So this gain level was chosen for enrolling the PD-100 vocabulary for each of the four test speakers.

The templates made at gain 3 were next tested at two signal-to-noise ratios with helicopter cockpit noise: +15 dB and +10 dB. At gain 3, the CSRDF recognizer responded frequently with its error message "stt 005", meaning "Capture Buffer Overflow". The user's manual describes this as indicating that the input utterance exceeded the maximum allowable template length of 1.8 sec. Apparently the background noise level was high enough to trigger audio input sampling by itself. Therefore, the recommendation of the vendor, which had been made for the VPC-2100, was used; the input gain was reduced during testing in noise to 2. At this gain level, the CSRDF recognizer, operating with maximum acceptance threshold, scored a nearly perfect recognition run on the numbers vocabulary for List 1 - correct recognition for 39 of the 40 numbers from one to forty - when the enrollment tokens for these words were used as the test tokens. When tokens from Enrollment List E3 were used (non-enrollment tokens) the percent recognition accuracy for the same 40 words was reduced to 67.5%. In order to eliminate false alarms from illegal tokens and insertions due to noise, it was decided to set the acceptance threshold to 27 for testing in noise, the same value recommended by the vendor for the VPC-2100. For testing in quiet, the vendor's default threshold of 50 was used.

Enrollment of PD-100 Vocabulary - The PD-100 vocabulary was enrolled for each of the four speakers at input gain 3. It was found that the shorter length of the PD-100 words, compared to the numbers vocabulary, made it possible to use two templates per word, so long as only one of the three Lists of PD-100 words was made active.

While input gain 3 was the best gain for most of the enrollment tokens, the input gain of the recognizer had to be changed for some words for some of the speakers. For speaker F2 it was necessary to re-enroll selected words at either gain 2 or gain 4, in response to status messages from the recognizer indicating that the enrollment tokens for these words had been spoken too loudly or too softly. For speaker M1, all but one token was successfully enrolled at input gain 3 and one token was enrolled at gain 4. For speaker M3, approximately 30% of the tokens had to be re-enrolled at the lower gain level 2. For speaker F4, about 20% of the tokens had to be re-enrolled at gain 2.

In order to determine if gain 2 would have been a better choice for speakers M3 and F4, tests were done of enrollment at gain 2. At this level, the tokens that had previously been successfully enrolled at gain 3 were reported by the recognizer as too low in amplitude. Thus, to get good templates for these speakers, it was necessary to adjust the input gain for each word individually.

After enrollment the templates for each speaker were verified in quiet conditions with enrollment tokens and with other tokens from the enrollment set that had not been used for enrollment to calibrate the performance of the recognizer.

Preparation Time

Preparation time for the CSRDF recognizer was short due to the use of the SCR software. All the necessary vocabulary, enrollment sequence, and syntax definition (set) files were created in 3 hours. The physical installation took an hour to install one of the CSRDB SCR computers in the PLRA laboratory and connect the serial line and the audio input line to the recognizer.

Template Enrollment Time

Template enrollment for the numbers vocabulary took 4.5 hours. Template enrollment for the PD-100 vocabulary for the four speakers took 30 hours. Most of the enrollment time was spent re-enrolling specific words at a different gain. The templates were calibrated simultaneously with the templates of the Smiths and of the Votan in order to save time. The calibration of templates for of all four speakers took 16 hours.

Problems

The greatest problem with the CSRDF recognizer was the apparently narrow dynamic range for input speech during enrollment. Speakers M3 and F4 do exhibit more variability in speaking level than do speakers M1 and F2. However, even for speakers M1 and F2 it was necessary to adjust the input gain for individual words. None of the four speakers exhibits the dynamic range that can be expected in normal flight operations, much less combat operations.

Another problem with the CSRDF recognizer is the limited information provided during connected recognition. The only output is the word number recognized and an internal event number, making it impossible to monitor goodness of fit of the user's speech with the templates.

Finally, the limited size of the active vocabulary memory constraints the number of words that can be active at a time. These constraints in turn limit attempts at designing command languages that use pilot's natural cockpit phraseology.

Useful Features

The system status messages during enrollment about the input signal level are very useful and would be even more useful were the dynamic range increased.

In the host mode of operation, the system is extremely flexible in the order of enrolling vocabulary and the context in which vocabulary is enrolled. This flexibility makes it easy to design enrollment procedures that will capture linguistically and phonetically representative tokens of the vocabulary with the coarticulation that can be expected during use.

APPENDIX C
PREPARATION OF THE ITT VRS 1280 FOR PD-100 TESTING

The ITT VRS-1280 is a single board system for the IBM-PC or AT buss. ITT supplies software to control the board under the xenix operating system for purposes of template enrollment and recognition. At the time of the PD-100 Phase III evaluation, ITT reported they had software to control the board under MS-DOS for recognition but that template enrollment software was available only under xenix. A procedure was available to then port the xenix-created templates to the MS-DOS environment for recognition, but this process was judged by PLRA as too involved for efficient testing. Therefore, ITT was permitted to supply the system for testing under xenix with the understanding the the Army would only be interested in a complete, stand-alone, MS-DOS version for compatibility with existing CSRDB software.

User's Manuals

Upon request from PLRA, ITT supplied two manuals, one for the VRS-2800, entitled "VRS-1280 Host Computer Software Documentation", and one for the xenix operating system, entitled "Xenix User's Reference". The VRS-1280 manual has sections on device driver installation, template generation, a speech digitizing and playback utility, a reference manual for the VRS recognizer and synthesizer (covering theory of operation, audio I/O, and computer I/O), and documentation for the VRS-1280 Application Development Library - a library of 'C'-callable subroutines to control I/O with the board, for file upload and download between the disk and the board, for file conversion from ASCII format to binary format for the recognizer, and for VRS board event and status detection (called notifier functions).

Physical Installation

The ITT VRS-1280 was delivered by the ITT technical representative, already installed in a Compaq Portable II computer running under the xenix operating system with two floppy disk drives and a hard disk drive. He had also written several custom C-shells to control ITT's enrollment and recognition programs in order to comply with the PD-100 procedures. He requested and was supplied with an oscilloscope, loaned by CSRDB, in order to monitor the input audio signal for distortion and overall level. The Compaq with VRS-1280 board installed was placed in the PLRA speech lab on a table beside the Smiths system.

Audio Interface

The VRS-1280 was jumpered for microphone level input since the CSRDF and CSRDB Development Laboratory systems use microphone level. Audio input to the device was via line 6 of the PLRA audio distribution system and was set to 10 mV RMS for a 1000 Hz tone presented at 94 dB SPL at the input microphone. At this level, the

ITT technical representative observed no clipping of the signal after it reached the VRS-1280 board.

Software for Enrollment

Software for enrollment was custom developed by the ITT technical representative using the 'C' shell feature to develop command script files that would automatically control the ITT-supplied "sbrhost" program - a very versatile, menu driven program for enrollment and recognition. For the enrollment, stock shells could be used as models for the most part. These shells required data files, as arguments, to define the vocabulary, and the order in which the words were to be enrolled. PLRA staff and the ITT representative each edited some of these data files, with the PLRA editing being done on an IBM-PC compatible and then read into xenix format by a xenix utility program. The data files were edited as ASCII files and then converted to binary format by an ITT supplied utility called "scriptcon". In all, 17 files - 2 shells and 15 data files were needed to enroll each of the vocabularies - 17 files for the practice vocabulary of numbers, and another 17 files for the PD-100 vocabulary. In addition, various template files and recognizer parameter files, supplied by ITT, had to be copied to the VRS-1280 memory and used as the basis for enrolling speaker-dependent templates. One of these is a set of templates, called SEEDS, for the digits zero through 9, supplied by ITT, for the appropriate sex speaker - male or female. Another set of templates, called "rejection filler templates" is generated automatically during the initial enrollment process for a particular speaker and must then be included for subsequent enrollment steps.

Software For Testing

Software for testing included a custom shell called "pd100", written the ITT technical representative, and a data file to define the vocabulary and syntax* for each of the three active lists of PD-100 words: 1a, 2a, and 3a. The "pd100" shell took as arguments file extension of the template files, the name of the binary version of the syntax file, and the directory and name of the file into which the test results were to be written. The data file was edited by PLRA staff as an ASCII file on an MS-DOS machine and then ported to the xenix operating system via the xenix file conversion utility. An ITT supplied conversion program

----

After reviewing the draft final report, ITT expressed concern that their recommended syntax for PD-100 testing had been designed to maximize recognition accuracy for isolated words with possible detrimental effects on connected word recognition. Had the test been conducted with a different syntax, the isolated word results might have been worse than, the same as, or better than those obtained and reported here. See Note 2, following the References, for details.

called "synwrite" was then used to convert the ASCII file into binary.

Noise Handling

The ITT system handles noise by sampling the background noise prior to recognition. The ITT technical representative created another custom shell called "calib" to perform this function. Calib was used for both the quiet (ambient room noise) condition as well as for the UH-60 noise condition. After observing insertion errors made by the system with the practice numbers vocabulary presented in the simulated helicopter noise, the technical representative proposed to make a template of the helicopter noise. He experimented with various durations and also tried several methods of adding this template to the syntax. Once he had a procedure that worked well, he made a shell to allow PLRA staff to create custom noise templates of the UH-60 noise. This shell first obtained a sample of the audio system noise - tape hiss, electrical noise. Then it captured a 10 second long sample of UH-60 noise played at the testing presentation level of 85 dB SPL. From this sample it made a template which if recognized by the system during testing would be identified as noise rather than as one of the PD-100 words.

Software for Data Collection

The custom "pd100" shell written by ITT captured the recognition results and wrote them to a file name specified by PLRA staff at run time. These files were in xenix format and were then converted by the xenix file conversion utility to MS-DOS format for PLRA data analysis.

Enrollment Procedure

The standard ITT enrollment process has four steps. First the ten digits must be spoken in several modes: as strings of 5 digits, then as strings of 3 digits, and then as single digits. Second, the digits are used as the first and last word in three-word phrases, with what ITT calls "carrier" words in the middle position. There are four carrier words and these are spoken in each of two such phrases, followed by ten more phrases in which the carrier words are in initial and final position and the digits are in medial position. Third, the actual vocabulary words are spoken in three-word phrases, using the carrier words in initial and final position and the vocabulary words in medial position. Each vocabulary word occurs in two such phrases. Thus 200 phrases were required for the 100-word PD-100 vocabulary. Fourth, and finally, each of the vocabulary words is spoken in isolation several times. ITT recommends more than two tokens for this. Under the terms of the PD-100 Evaluation, however they were limited to two isolated tokens.

Additionally, a compromise was reached concerning the other steps in the enrollment process. Since the other four recognizers were limited to just two isolated word tokens, it is not scientifically valid for recognizer comparison purposes to allow the ITT system to have, in addition to these, two connected word tokens and the various digit and carrier word tokens. On the other hand, ITT did believed their system's performance would not be representative unless their enrollment procedure were followed. It was therefore agreed that two of the PD-100 speakers - a male and a female - would be enrolled using the ITT four-step procedure but with a limit of two isolated tokens per word. The other two speakers - a male and a female - would be enrolled using the templates of the first two speakers as seeds or starting templates and modifying them only with two isolated tokens per vocabulary word spoken by these second two speakers. In any event, ITT understood that any new CSRDF recognizer will have to be able to perform well with a maximum of only 2 isolated templates per word.

Preparation Time

Preparation time included 6.5 hours for hardware checkout, setting of input audio levels, and 3 hours to edit the data files needed to enroll the practice numbers vocabulary. The second day, 13 hours, was spent in an unsuccessful attempt to enroll the numbers vocabulary using seeds which ITT had prepared prior to arrival but which had been made by a different audio delivery system with different frequency response characteristics compared to the PLRA audio distribution system.

Four hours were required to edit the necessary data files for enrolling the PD-100 vocabulary. This was done by PLRA staff while the ITT technical representative during the same four hours continued to create custom 'C' shells for the enrollment and testing phases.

In all, 30.5 hours were spent in preparation.

Template Enrollment Time

The third day 13 hours were spent making the templates for the numbers vocabulary using the full ITT procedure. Some time was spent determining by trial and error the best duration for what ITT calls the silence template for the particular female speaker who had recorded the numbers vocabulary. Additionally, recognition performance for the numbers was better when the templates had been made in the context of a sample of the low level background hiss of the audio tape that contained the numbers tokens for enrollment.

Enrollment of PD-100 words in phrases for F2 took only 2 hours, since carrier words and digits had already been created for this speaker during the numbers vocabulary enrollment. Another 6.5 hours were needed for isolated word enrollment, verification, re-enrollment of selected words, and calibration for speaker F2.

Enrollment of PD-100 words in phrases for M1 took 3 hours including enrollment of digits and carrier words. Enrollment of PD-100 words in isolation took another 5.5 hours, including verification, re-enrollment, and calibration.

For speaker M3 the templates of PD-100 words in phrases from speaker M1 were used as the seeds, and only two isolated word tokens per vocabulary word spoken by speaker M3 were used for enrollment. This process took 5 hours for enrollment and re-enrollment of problem words, followed by template calibration.

For speaker F4, 3 hours were initially expended unsuccessfully trying to enroll isolated words using speaker F2's templates in phrases as the seeds. The problem was false triggers of the enrollment algorithm and was eventually traced to very low volume "ghost" copies of the enrollment tokens on the token tape, caused by bleed through of the magnetic field of the actual recording of those tokens. The ITT system was the only system so sensitive as to trigger on these very low level signals. These faint copies were so low level as to be nearly undetectable a human listener with headphones concentrating very hard on listening for them. An attempt was made subsequently to measure their level relative to that of the actual tokens, but they did not exceed the background level of the tape hiss. Nevertheless, both PLRA staff and the ITT technical representative were convinced that the system was triggering on these low level copies of tokens after a blind listening test in which a PLRA listener noted when these ghost copies occurred while not looking at the recognizer screen and the ITT technical representative noted false triggerings without having the headset on for listening. A nearly perfect correlation was found between the two sets of observations. And, the problem was solved by advancing the tape past the ghost tokens and beginning playback at the onset of the actual tokens.

Isolated template enrollment was restarted for F4 and took 5.5 hours, including verification, re-enrollment, and calibration.

Development of the noise template, including testing of several alternative designs, took 9 hours.

Completion of enrollment of the ITT system was delayed for two weeks because of a hard disk crash which required that the Compaq computer be shipped back to ITT for rebuilding of the hard disk. Fortunately only the boot tracks of the disk were affected and none of the template files already made were lost.

In all, enrollment of tokens for the four speakers took 52.5 hours.

Problems

The lengthy enrollment procedure and the sensitivity of the

ITT system to very low volume signals resulted in an unacceptably large expenditure of time. The sensitivity of the system to low volume input was also a problem during testing. Normally, all recognizers are left in recognition mode throughout a list of PD-100 words or phrases. When the experimenters talk in a low volume voice to compare their observations for a particular word, the recognizers do not respond. However, the ITT system almost always responded with one or more words recognized, albeit not always accepted as legal, when the experimenters conversed between PD-100 test tokens. This problem was so frequent that it was decided to take the ITT out of recognition mode after its response to each PD-100 word. Thus, the insertion error rate for the ITT is unrealistically low compared to what it would have been had the system been treated like the other four recognizers and left in recognition mode throughout a particular test token list.

Another problem for rapid development and testing of a vocabulary is the large number of different files that must be created in order to enroll and test a vocabulary. Without the help of the ITT technical representative to create shells and many of the vocabulary and syntax files for enrollment and testing, the preparation time would have been much longer.

Finally, the template enrollment process was susceptible to unnoticed errors which when undetected seemed to propagate throughout the template set. This seems to be because a given template is formed or modified on the basis of already existing templates. If a bad template is present, then there is a risk of its contaminating subsequent templates as they are made. The ITT supplied software does provide an indication that a template may not be very good. It does this by reporting that the match between an existing template and the new token is not close enough to be automatically accepted. A user unfamiliar with the behavior of the system is likely to manually accept the new token and unwittingly contaminate the templates. This actually occurred during the PD-100 enrollment and was corrected only because the ITT technical representative recognized the symptoms was able to judge when to delete the templates and begin again.

Until ITT releases an MS-DOS version of their software, it will not be possible to be certain of the ease or difficulty of interfacing the VRS-1280 to the CSRDB SCR. The menu driven enrollment and recognition program and the shells would probably not be usable for this purpose. To the extent that the application library 'C' functions could be used for control of the device, the interfacing would be made easier. But, even with these functions, a major software modification to the SCR would be needed.

The lack of any built-in feedback on audio signal input level results in a lot of trial and error during enrollment to find the best input level. Expecting a user to open up the computer and put an oscilloscope on the pin of an integrated circuit to look at the audio waveform is unrealistic.

The extreme sensitivity of the system to having a sample of exactly the noise spectrum and level of the background noise for both enrollment and recognition results in wide variations in recognition accuracy. Several times during the enrollment process, a change in background noise sample produced dramatic differences in enrollment success. It is conceivable that this sensitivity was in part responsible for the significantly poorer performance of the ITT system in noise compared to the CSRDF recognizer.

Useful Features

Unlike any of the other recognizers, the ITT system makes a separate file for the template of each word. Thus a library of templates is possible, and application vocabularies can be made up from the library. It should be possible to merge templates from several speakers to create group-dependent template sets for use by small groups of users. Most other recognizers store all the templates for a particular vocabulary in a single file, and this file may or may not be easily editable to merge in the templates from another file.

The extreme dynamic range of the audio input, while a problem in open recognition, may well be an asset when a manual push-to-talk switch is used to control audio input to the recognizer. Certainly a weakness of most recognizers is their limited dynamic range in comparison to the normal dynamic range of human speech, not to mention the range that occurs in flight for routine operations compared to emergencies or combat.

The rejection filler templates of the ITT system seem to have been designed to take care of false alarms and often performed that function. While the system often attempted to recognize one of the illegal test tokens as a legal word, a rejection filler template usually formed a better match than any of the legal words and thus resulted in a correct rejection.

APPENDIX D
PREPARATION OF THE MARCONI MACROSPEAK FOR PD-100 TESTING

The Marconi Macrospeak is a stand-alone unit, 210 mm x 380 mm x 85 mm (12.2" x 15.1" x 3.7"), weighing 7.3 kg (16 lb), powered by either 110-115 V. or 230 V. It has two RS-232 serial ports with programmable baud rate - a VDU (Visual Display Unit) port for a terminal and a host port with hardware and XON/XOFF handshaking. A manual foot switch is supplied to toggle the unit in and out of recognition mode. This toggle can also be controlled via software. The unit is supplied with a Shure SM-10A lightweight unidirectional dynamic boom microphone mounted on a headband. The Macrospeak has a built-in 3 1/2" floppy disk drive for saving and retrieving a file containing vocabulary, templates, syntax, and macros (word-specific subroutines that are to be executed when the associated word is recognized). One file per disk is permitted. The vocabulary, templates, syntax, and macros file can also be up- and down-loaded via the host port.

The Macrospeak is also available as a set of two boards, without the internal 3.5" disk drive, power supply, or enclosing box. The boards do include ROMS with the macros and the terminal and host interface software.

User's Manuals

Marconi provided four manuals with the Macrospeak. These were

1) "An Introduction to Macrospeak, January 87, Issue A",

2) "Macrospeak User Handbook (Draft Only) July 86",

3) "Speech Systems Division Technical Report (Draft) Special Software Features, June 87", and

4) "Speech and Information Systems Division Technical Memo: Wildcard, 15 March 1988".

The third and fourth manuals were supplied under a proprietary non-disclosure agreement between Marconi and PLRA.

The "Macrospeak User Handbook" covered nearly all the information that was needed to prepare the Macrospeak for the PD-100 testing. For simple operations with a terminal, the Macrospeak's menus provided easy prompts for control of the system once a vocabulary, optional syntax, and optional macros had been entered. Template enrollment, setting of a noise mask, setting of input gain, and switching between set-up mode (for vocabulary and syntax editing, enrollment, etc.) and recognition mode could be done almost without reference to the manual. The menu control software was robust, did not blow up when an illegal entry was made, and provided good visual feedback to the user regarding state of the system. However, for creating syntax and the macros (similar to subroutines) to be executed for each different word recognized, a background in assembly language programming and assembler directives turned out to be extremely useful in understanding the manual sections on these operations. The information in the manuals was greatly augmented by excellent technical support from the Marconi technical representative.

The "Introduction to Macrospeak" gave a good overview of the system operation and features but did not really explain how to use it. The other two technical manuals provided a wealth of information that would be useful to the serious user who is prepared to invest time in application software development. They described a number of useful engineering features which could be used to fine-tune the system for idiosyncrasies of individual pronunciation and noise backgrounds, to display template data, to examine the input to the audio front end, and to suppress the menus for faster and more efficient program control of the Macrospeak directly via its host port.

Physical Installation

The Macrospeak was placed on a laboratory table together with an IBM-PC compatible computer also provided by Marconi for the evaluation. The Macrospeak sat on top of the computer box with disk drives, and the monitor for the computer sat on top of the Macrospeak. The VDU port of the Macrospeak was connected to the COM1 serial port of the PC. A terminal emulation and communications program, Procomm by PIL Software Systems, was used to send commands to the Marconi and to capture and save its recognition responses to an ASCII text file. The host port was also tested and was used to download the vocabulary, syntax, and macros for the PD-100 testing after editing these with a text editor on the PC.

Audio Interface

The Macrospeak was connected to the PLRA audio distribution system line 2. The audio signal level was set to 1 mV RMS for a 1000 Hz tone, a level which was mid range between the recommended 0.5 to 2 mV RMS for microphone input to the Macrospeak.

Software for Enrollment

Software for enrollment consisted of text files, prepared according to the format specified in the user's manual, to define the vocabulary text strings, to assign vocabulary words by number to sets, and to define the macros that would be executed upon recognition of each vocabulary word. These files were edited using a text editor on a PLRA IBM-PC/XT compatible and then downloaded to the Macrospeak via its host port and using download commands from its menu-driven operating system. Two files were required, one for the numbers vocabulary, used for set-up and checkout prior to enrolling the PD-100 words, and one for the PD-100 vocabulary. The syntax conformed to recommendations given by the technical representative in collaboration with the Macrospeak engineering design staff.

Software for Testing

No additional software was needed for testing beyond the two vocabulary, syntax, and macro definition files that were created for enrollment.

Noise Handling

The Macrospeak has a feature called the "noise mask". The system, upon command via a menu item, samples the ambient noise and then uses this sample when recognizing speech spoken in noise. A trial noise mask was created for the simulated helicopter noise that was used for set-up and checkout with the numbers vocabulary. A second noise mask was created for the UH-60 helicopter noise at the 85 dB SPL test level. This mask was used for all PD-100 test runs in noise.

Software for Data Collection

Software for data collection consisted of an off-the-shelf terminal emulation and serial communications program, Procomm, by PIL Software Systems, a copy of which is owned by PLRA. Procomm's text capture mode was activated whenever the Macrospeak was in recognition mode for testing. The macros which had previously been programmed for each of the vocabulary words caused the Macrospeak to display on the monitor the word number, text string, and score of each word that was recognized. Since the output to the monitor, via the Macrospeak VDU port, was being captured by the Procomm terminal emulator, this text could then be saved in an MS-DOS text file by Procomm. The macro for the wild card was written to cause an asterisk to be displayed each time the wild card was recognized. These wild card indicators were thus included in the data file allowing an assessment of the level of activity of the wild card feature.

Enrollment Procedure

The Macrospeak required one token per word, spoken in isolation, for making templates. With one template per word, it performed continuous, speaker-dependent connected word recognition.

Using the numbers vocabulary, templates were made at three different input gains and each set of 100 templates was tested at the same gain at which they had been enrolled. The Macrospeak has a built-in function to select the gain when the word "Macrospeak" is spoken to it. Using a presentation level of 100 dB SPL for several recorded tokens of this word, the system selected a gain of 4 more often than any other value. Templates were thus made and tested at gain 4. Templates were also made and tested at gain 3 with the result that there were a small number of misses. Finally, templates were made and tested at gain 5. Recognition accuracy at gain 5 was the same as for gain 4, i.e. 99% for the enrollment tokens, but the distance scores at gain 5 were slightly larger. Therefore, it was decided in consultation with the technical representative to use a gain of 4.

The PD-100 words were then enrolled at gain 4 and calibrated. The tokens used for calibration were the enrollment tokens and a second set of tokens that were available as alternate enrollment tokens but which had not been used for enrollment of the Macrospeak. Of all one hundred words, only those words in Set 1 of the PD-100 were made active. However, all PD-100 words were presented to the recognizer. The distance scores thus obtained provided raw data which was used to select a rejection threshold for use in the PD-100 testing. Based on an inspection of the calibration data for each of the four speakers, the technical representative selected a value of 9.

A calibration run was also done on the numbers vocabulary, spoken by speaker number 2, with the simulated helicopter noise, using the appropriate noise mask. Based on these data, the technical representative decided to also use a threshold of 9 for the PD-100 test noise runs.

The numbers templates, with vocabulary, syntax, and macros, for speaker 2 were saved on 3 1/2" disk as were the PD-100 templates for each of the 4 speakers.

Preparation Time

Preparation time included reading the manuals, setting up the system to communicate with the PC, learning the macro language, creating and testing the enrollment files for the numbers vocabulary and the PD-100 vocabulary, and setting the appropriate input audio signal level for the unit. 57 hours were expended in preparation.

Template Enrollment Time

Template enrollment time included enrollment and calibration of the trial numbers vocabulary for speaker 2, the enrollment and calibration of the PD-100 vocabulary for all four speakers, and generation and testing of the noise mask. 16.5 hours were expended in enrollment and calibration. Actual enrollment of 100 words was accomplished in about 15 min. per speaker. The rest of the time was spent calibrating, printing out and analyzing the data.

Problems

There were few problems with the Macrospeak. One annoying problem was the lack of a POWER ON indication. The only ways to determine if the unit was on were to listen for its very quiet cooling fan or to feel the position of the power switch and try to remember which was the ON position. The menu-driven operating system at first appeared to be a problem for interfacing to a computer, e.g. the CSRDB Smart Command Recognizer. However, the expert mode, which toggles off the menus, should remove that as a problem. The limitation of only one file per 3 1/2" disk is somewhat expensive in terms of the number of disks and the associated storage space required to keep a library of templates for different users and different speech command sets. Presumably the up and downloading of files to and from the SCR would be done only during command set development and template enrollment. At run time it would be preferable to have templates on the 3 1/2" disks for individual pilots.

The manuals were difficult to understand, even for a somewhat experienced assembly language programmer. Trial and error had to be used in order to figure out how to write macros and to define the syntax.

Useful Features

The Macrospeak has a number of features that would be useful for the CSRDB R&D program. The noise mask allows enrollment in quiet and recognition in a variety of noise backgrounds and levels, using the same original template set. A confusion matrix can be generated for any ten user-selected words and used to discover word pairs which are highly confusable to the recognizer. This information can be used to select words that should be re-enrolled. The wild card enhances rejection accuracy and can probably be used to make the Macrospeak perform in a word spotting mode. The macro language is itself rather powerful, includes conditional branching, and permits setting the rejection threshold individually for each word, if desired. Also, during download of templates, vocabulary, syntax, and macros, each of these four types of data can be independently downloaded without affecting the other data types already in the Macrospeak memory. There thus appear to be no insurmountable barriers to interfacing the Macrospeak to the SCR.

APPENDIX E
PREPARATION OF THE SMITHS SIR-L FOR PD-100 TESTING

Smiths Industries supplied their SIR-L as a stand-alone system, incorporated into a microcomputer with the PDOS operating system and a VME buss. The PDOS operating system bears some resemblance to DEC operating systems. The unit is supplied with a Tektronics 4107 color graphics terminal which is used in alphanumeric monochrome mode or full color graphics mode, as appropriate. The recognition software is menu driven and provides the following capabilities: vocabulary file creation and editing, syntax structure creation and editing, editing of the audio input buffer, creation of dynamic time warping (DTW) algorithm templates, creation of hidden Markov model templates, creation of neural network templates, editing of any of these templates using a three dimensional color graphic display of the template, mixing of templates of all three algorithms for a given recognition vocabulary, an active vocabulary of 96 words, vendor-reported 120 ms. response time from end of word to recognition response, sampling of background noise for development of a noise mask to use during recognition, and recognition trace analysis - a slower-than-real-time color graphic display of the recognition matching process in action for the DTW and the hidden Markov templates. All the above capabilities were exercised to greater or lesser degree by PLRA during the PD-100 Phase III evaluation. The recognition response time was not measured; however it was observed to be faster than that of the CSRDF device. The DTW algorithm was selected for evaluation because it was the only one that could meet the restriction of no more then two tokens per word for enrollment. However, the Smiths technical representative informally demonstrated that templates made with the hidden Markov modeling algorithm and the neural network algorithm could perform accurately over a wider dynamic range of audio input, including shouting, than could those made with DTW algorithm.

The Smiths system as supplied for this evaluation is closer to a pre-production prototype than to a production system. The engineer who came from England to serve as the technical representative made a number of changes to the system software during the 4.5 days which he spent at PLRA in order to accommodate the PD-100 vocabulary size and syntax requirements. He was handicapped by not having taken delivery of the latest version of the PDOS operating system prior to bringing the system to PLRA for evaluation.

USER'S MANUALS

The documentation supplied with the SIR-L consisted of a 24-page technical note, entitled "Technical Notes on Smiths Industries 'SIR' Speech Recognition Equipment", dated 24 June, 1988. As such the documentation was high level, more like a theory of operation than a user's manual. Verbal instruction from the technical representative was essential to learning how to operate the system. This was supplemented by careful trial and error on dummy template, vocabulary, and syntax files.

Physical Installation

Physical installation of the system was done by the technical representative and consisted of placing the unit on one of the PLRA lab tables, connecting the Tek terminal via RS-232 cable, and connecting the audio input cable supplied by PLRA to the system audio input.

Audio Interface

Audio input was supplied via line 5 of the PLRA audio distribution system. A 1000 Hz tone presented to the input microphone at 94 dB SPL resulted in a signal level of 3.6 mV RMS input to the SIR-L.

Software for Enrollment

The SIR-L recognition program menu provided an automatic enrollment procedure that enrolled DTW templates for each word in the vocabulary in order of word number. Prior to enrollment, vocabulary files had to be created, using a menu-driven vocabulary file creation and editing option from the recognition program. The menu-driven nature of this editor made vocabulary file creation a lengthy process, and modifications were time-consuming compared to editing an ascii file with a screen editor.

Software for Testing

For testing it was also necessary to create syntax files. These were created using another menu-driven editor, itself a selection from the recognition program main menu. The creation and editing of syntax via menu was even more laborious than that of vocabulary files.

Noise Handling

The system automatically takes a sample of the background noise when it enters recognition mode from the main menu of the recognition program. When tests were performed using the practice numbers vocabulary in +15 and in +10 dB S/N for the simulated piston engine, 2-blade rotor helicopter noise, it was found that the acceptance threshold had to be set relatively more open to obtain recognition responses compared to the quiet condition. The technical representative selected a value of 12 for testing in quiet and increased this to 15 for testing in noise. Without the more lenient threshold, the system gave a large percentage of miss responses to legal words when the background noise was present.

In addition to adjusting the threshold, a second technique was tried for noise handling. One of the menu options allowed for creating a template from any time slice of the input audio buffer. At the suggestion of the technical representative, a template of the simulated helicopter noise was made and added to a copy of the templates of speaker F4. The vocabulary and syntax files were edited to include this "word" template. However, when recognition was tried, the SIR-L constantly recognized the noise template to the exclusion of all other words. Therefore, the noise template was not used for actual PD-100 testing.

Software for Data Collection

Software for data collection did not exist as such. Rather, it was planned to interface the SIR-L via its terminal serial port with a Y-connection to a second microprocessor that was in terminal emulation mode. This microprocessor would capture all text output from the SIR-L and store it in an MS-DOS format ASCII data file. Baud rate, stop bits, and word length of the terminal emulation program were set to match that of the SIR-L. However, after several unsuccessful attempts to implement the serial interface, this plan was abandoned, and a hand-written copy of what appeared on the Smiths terminal was made by one of the Experimenters during the data runs.

Later, just prior to returning the SIR-L to Smiths, one last attempt was made to implement the serial data transfer. It was found that the Tek terminal had to be disconnected and complete control passed to the computer that was running the terminal emulation program. Also, the terminal emulation program had to be configured to emulate a DEC VT-100 terminal using XON-XOFF handshaking protocol. Then, so long as only text was being output by the SIR-L, data capture was successful.

Enrollment Procedure

The enrollment procedure was simple with few options. One token per vocabulary word was required. The vendor-supplied program prompted for each word in order of word number, starting with word number 1 and continuing to the end of the vocabulary as defined in the vocabulary file.

Preparation Time

Preparation time included modifications made to the software by the technical representative at PLRA to try to accommodate the PD-100 maximum active vocabulary size of 100 words, to display the recognizer output on the terminal screen for data collection, and to add some functions to the the menu-driven editing of the syntax. He spent about 15 hours during the first 2.5 days with these modifications. Seven hours were spent by PLRA the first day in system familiarization and in instruction provided by the Smiths technical representative along with an initial test of the practice numbers vocabulary using a file edited by the technical representative. Another 8.5 hours were spent by PLRA creating and editing vocabulary and syntax files for the PD-100 and numbers vocabularies.

Template Enrollment Time

Template enrollment took relatively little time. The numbers vocabulary was enrolled, verified, and calibrated in 45 minutes. Then, since it had not been possible to increase the active vocabulary size to 100 words, due to constraints resulting from the older version of the operating system, the technical representative edited the numbers vocabulary template file into two files, one containing the templates for PD-100 Lists 1 and 2, and the second containing the templates for PD-100 List 3. This editing and software modifications to restore the 96 word maximum active vocabulary took another 4.5 hours. After the division of the vocabulary, calibration of the numbers vocabulary templates took 4.5 hours. One reason the calibration took so long was that all calibration data had to be hand-written due to the lack of a method for collecting these data in a file.

Templates for all four PD-100 speakers were enrolled in 3 hours. Calibration for one speaker, M1, took an additional 3 hours. The templates of the other three speakers were calibrated at a later date, in parallel with those of the Votan and CSRDF recognizers, to save time. In parallel, another nine hours were thus spent with the template calibration.

Three hours were spent testing the numbers vocabulary in the presence of the simulated helicopter noise at +15 and +10 S/N and selecting the acceptance threshold of value 15 to use for the PD-100 data collection in noise.

In all, 27.75 hours were spent with template enrollment, testing, and calibration for the Smiths SIR-L.

Problems

File editing via menu was time-consuming and error prone. The syntax was constrained such that recognition of any word in a particular set forced a change of set to the same set regardless of which word had been recognized. Natural language is not so constrained, and the CSRDB SCR allows specification of a different set of following words for each individual word in the current active vocabulary. The SCR syntax interpreter also permits a particular word to be used in different locations in the command phrase, depending on the syntactic function of the word for each different command phrase. The Smiths syntax forced a particular word in a particular set to always occupy the same position in the command phrase.

Useful Features

The SIR-L provided a number of very useful features for the research and development of speech command systems.

Audio input levels were displayed on an LED display on the front of the unit, allowing easy adjustment of input gain and providing a user with immediate feedback on his or her speaking level.

The ability to see a three-dimensional (3D) representation of each template and to edit it is extremely useful for testing hypotheses about the characteristics of normally spoken speech compared to shouted speech, speech spoken under stress or fatigue, and for comparing the speech of "sheep" to that of "goats".

In conjunction with the 3D template display, the dynamic display of the recognition process (in slower-than-real-time) permits further testing of hypotheses about why recognition is degraded for speech spoken under stress, etc.

The option to include some Markov model and some neural net templates, for words that exhibit greater variability, together with conventional DTW templates offers the potential for a template set that will be more robust in the face of intra-speaker variability than are template sets created by other recognizers using only the DTW algorithm. The drawback to Markov models and neural nets is that template creation requires a large number of tokens and several hours of off-line processing. This is not practical for normal CSRDF operations. However, the making of Markov model or neural net templates for only a few words that exhibited poor recognition might be acceptable in the simulation schedule. Further experimentation, however, would be required to determine whether there are indeed performance advantages to this hybrid template approach and to determine the time and cost trade-offs.

APPENDIX F
PREPARATION OF THE VOTAN VPC-2100 FOR PD-100 TESTING

Votan supplied a Model VPC-2100 speech recognizer board which plugs into a standard IBM-PC compatible buss.

User's Manuals

Two manuals were supplied with the VPC-2100. Voice Key, and Voice Library.

Physical Installation

The VPC-2100 was installed in an IBM-PC/XT compatible microcomputer. It's default configuration uses the COM2 as I/O for the recognizer. In order to test the board in the configuration in which it would be used for the CSRDF, this default had to be changed to COM1. The SCR uses COM1 for I/O with its speech recognizer and uses COM2 (the VPC-2100 default) for a speech synthesizer when in stand-alone mode and for communications with the host computer in host mode. The conversion to COM1 involved both hardware and software. We had to change a jumper on the board and modify and recompile the VPC-2100 software module, written in the C language, responsible for I/O with the board. The Microsoft C Compiler, version 5.0 or higher, using the large model, was required for this recompilation. PLRA's licensed copy of version 5.1 was used for this operation. The PC with the VPC-2100 installed was placed on a laboratory table next to other units under test.

Audio Interface

Audio input to the VPC-2100 was from line 4 of the Audio Distribution System at a level of approximately 1 mV RMS (1.3 mV RMS for a 1000 Hz tone at 94 dB presentation level).

Software for Enrollment

Votan supplied their "Voice Keys" software, designed to let a user control a keyboard with spoken commands. They also supplied their "Voice Library", for use in developing custom speech recognition programs. Since the Voice Keys software permits a maximum of 64 words, it was necessary to write a custom program to handle the 100 words in the PD-100 test. Votan supplied on disk several example programs which used their Voice Library C functions to perform isolated and connected enrollment. None of these programs met our requirements for changing vocabulary and syntax and templates independently, nor for the different syntax node sizes required by the PD-100 test. The source code for one of the example programs, "sstest.c" was selected as a starting point and was modified and expanded to include the other capabilities needed for the PD-100 enrollment and recognition testing. We successfully incorporated several of the SCR modules for compatibility with SCR vocabulary file format and also had to write some special purpose functions. The resulting custom program, which we called "vpctest", was a menu-driven program which gave the experimenter options to to the following operations: 1) Display current input gain and acceptance threshold, 2) Set a new input gain, 3) Set a new threshold, 4) Read in a vocabulary file, 5) Read in a Set file, 6) Enroll all words in the vocabulary, 7) Enroll a particular word number, 8) delete all templates for a particular word number, 9) Test recognition and store the results in a data file, and 10) Exit.

In addition to writing the vpctest program, we prepared text vocabulary files containing the text strings of the numbers vocabulary and those of the the PD-100 vocabulary. We also prepared text set files to assign words by vocabulary word number to sets for List 1, List 2, and List 3 for template calibration.

Software for Testing

The "vpctest" program was also used for testing. Text files were created to assign PD-100 words to subsets List 1a, List 2a, and List 3a to form the legal word sets for PD-100 testing. The Test Mode of the program included data collection and saving to an MS-DOS format text file.

Noise Handling

Background cockpit noise during testing was handled by two methods which were used in parallel: 1) threshold limit and 2) gain limit.

Threshold Limit Method. The numbers vocabulary, which was used for set-up and checkout, was tested in simulated helicopter noise at signal-to-noise ratios of +15 and +10 dB. Tests were run at different acceptance thresholds in the range of 25 to 70 to determine the threshold which was needed to eliminate false alarms, actually insertion errors due to the noise, by the recognizer. It was found that insertions occurred for any threshold above 27, for both signal-to-noise ratios. At threshold 25, the recognizer missed recognizing one of the tokens that had been used to enroll it. Therefore, in consultation with Votan's technical staff, the threshold value of 27 was selected for the tests in noise.

Gain Limit Method. The second method of handling noise was recommended by Votan's technical staff. The testing in noise was done with the recognizer audio input gain set to one level lower than the gain at which the templates had been enrolled. This was in contrast to the testing in quiet conditions for which the input gain was the same as that used for enrollment.

Software for Data Collection

Software for data collection was written as part of the custom "vpctest" program. Nearly all the modules were taken, without alteration, from the SCR modules for data collection. The vpctest program asked the experimenter for a data file name, prompted for a header line, and then wrote data for each recognition event into that file during testing. Data collected included word number recognized, text string for that word, distance score for that word, template number recognized, second closest matching word number, distance score for second best word, and template number for second best word.

Enrollment Procedure

Checkout of the vpctest program, vocabulary files, and set files was done with the numbers vocabulary. Templates were made at input gain levels of 2, 3, 4, and 5. The best recognition was obtained for templates made at gain level 4, so this level was used for enrolling the PD-100 vocabulary.

The PD-100 vocabulary was enrolled using two tokens per word. In discussions with Votan marketing and technical staff, we learned that they sometimes recommend three or four templates per word for some words, but they accepted the constraint that we have for the CSRDF recognizer of a maximum of two tokens per word, due to limited time available to enroll individual Army pilots who serve as test pilots in the CSRDF and in the CSRDB Development Station.

For each speaker, templates were made using the tokens from the first two enrollment lists. These templates were then tested on the tokens in the first enrollment list and on the tokens in the third enrollment list. Words for which neither the first nor the second choice was correct were targeted for re-enrollment. Usually the enrollment tokens (those from the first enrollment list) were recognized correctly and the errors occurred for the non-enrollment tokens (from the third enrollment list). A copy of the templates was made, the original set saved, and the copy then modified. All templates for the words to be re-enrolled were deleted and these words were then re-enrolled from enrollment lists three and four. The test using tokens from enrollment lists 1 and 3 was then repeated. The template set which gave the best performance was then selected for use in the PD-100 testing. This procedure resulted in the original template set being used for speakers M1 and F4 and the modified set being used for speakers F2 and M3. In the case of speakers M3 and F4, the input gain level 4 during enrollment resulted in some problems. These speakers exhibited relatively more variability in speaking level than did speakers M1 and F2. Several of their enrollment tokens were 3 to 5 dB above the 100 db presentation level used for enrollment.

For speaker M3, it was not possible to store two templates per word in available memory due to the relatively larger amount of memory needed to store templates for those words which were spoken loudly. Accordingly, one template was made for each word at gain 4 and the second template was made at gain 3. This resulted in a template file of size 43.6K. These "hybrid-gain" templates were then tested at gain 3 and at gain 4. The best performance was obtained at gain 4, so this gain was used for PD-100 testing.

For speaker F4, tokens made at input gain level 4 were tested at gain 3 and at gain 4 and tokens made at gain 3 were tested at gain 3. The best performance was obtained with templates enrolled and tested at gain 4, so these were used for PD-100 testing.

Preparation Time

Preparation time, including board installation, software driver modifications, development of the vpctest program, and editing the vocabulary and set files for enrollment and testing, was 88.5 hours for PLRA staff. In addition, Votan supplied the services of one of their top programmers during a period of 2 weeks. By far the largest number of hours was required for developing the custom vpctest program to provide the basic functions needed for enrollment and testing of a 100-word vocabulary. The breakdown is given below:

Review VPC-2100 installation and programming documentation	2 hours
Board Installation	1 hour
Driver Modification (including 13.5 hours of unsuccessful search of documentation for interrupt number to change)	14.5 hours
Development of vpctest program. Partial development by PLRA - Menu, Read Vocabulary File, Read Set File, Enroll Templates, Test Recognition, Save Recognition Data to Disk, List Vocabulary. Unable to make Voice Lib functions work for saving templates on disk.	52.5 hours
Vpctest program source code given to Votan programmer for debug and addition of remaining functions: Delete All Templates, Delete Templates for One Word, Enroll Templates for One Word, Save Template file, Restore Template file from disk.	Calendar time: 2 weeks
Install and test Voice Library	

upgrade provided by Votan. Install and test vpctest program returned from Votan. PLRA Add remaining functions: Display Gain and Acceptance, Change Gain, Change Acceptance, Write Gain and Acceptance level to Data File.	16.5 hours
Edit vocabulary files and set files for enrollment and testing of numbers vocabulary and of PD-100 vocabulary.	2 hours
TOTAL PLRA TIME	88.5 HOURS

Template Enrollment Time

Template enrollment time for the numbers vocabulary, including testing at different input gains and testing different thresholds in noise took 15.5 hours. Votan marketing and technical staff observed 7.5 hours of the numbers vocabulary enrollment, including the enrollment and testing in noise at different input gains.

Template enrollment of the PD-100 words for the four speakers took 19 hours. Calibration of the VPC-2100 templates was done by PLRA staff in conjunction with calibration of the templates for the CSRDF recognizer and for the Smiths recognizer, to save time. Calibration took 4 hours per speaker for a total of 16 hours.

Problems

By far the worst problem with using the VPC-2100 was the extensive effort needed to develop a custom C-program in order to enroll and save templates, test recognition, and save the recognition data on disk, for a vocabulary of 100 words, while varying under program control basic parameters of input gain, acceptance threshold and assignment of words to the active recognition set. This effort took far more time than had been allocated for preparation of the device for testing. The functions in the Voice Library provide the building blocks for extremely versatile and powerful custom software, but a programmer experienced in C programming and in speech recognition is essential to take advantage of the capabilities of the Voice Library. In contrast, the CSRDF recognizer, an earlier model by the same vendor, provides higher level functions in the host computer mode, making it less demanding of programmer time to develop software for controlling the recognizer to perform the type of operations needed for the PD-100 test. For example, the CSRDF recognizer (a Votan 6050) assigns words to a set with a single line of code. In contrast, the VPC-2100 using the

Voice Library required 150 lines of c code for the function called (readset()).

The documentation provided detailed descriptions of each of the Voice Library functions, but these were grouped in three different chapters, apparently by chronological development. The user thus had to search each of the three chapters for any given function. One group, in alphabetical order by function name, cross-referenced by type of function, would have made a far more efficient reference.

Information on the software modifications needed to configure the VPC-2100 for any other than the default interrupts was not available in any of the documentation. The user was not even informed that a software change was needed. The author finally figured out the correct interrupt by studying the listing of one of Votan's sample programs, together with a reference book on the MS-DOS system software. The actual information, the search for which took 13.5 hours, could have been put in a five-line table and explained with a short (ca 10-line) paragraph.

Useful Features

The VPC-2100 reports not only the best match word for each incomming audio event, it also reports the second best match. In cases of substitution errors for the best match, the second best match is often correct. This information is useful to a program developer for automatic error correction and for assessing the user's speech variability, among other uses.

The VPC-2100 also reports which of the possibly several templates for a word provided the best match. The information is useful in similar ways as the reporting of second best match. It also can be used to determine which templates are the most useful and which, if any, are not used and could be eliminated.

In contrast to the CSRDF recognizer, which provides neither of these data types during connected recognition and provides the second best word match only in isolated recognition, the VPC-2100 is more informative and provides more useful information to the developer and user about the recognition events.

The inclusion of source code for sample C programs which used the Voice Library functions is a very helpful aid to the documentation and would be even more useful if the code itself were better documented (this author does not subscribe to the claim that C-code is self-documenting).

During enrollment of the practice vocabulary and of the

PD-100 vocabulary, it was observed that the VPC-2100 exhibited a dynamic range for audio input that was greater than that of the CSRDF recognizer. For the same recorded tokens presented to the two recognizers, fewer of these had to be re-enrolled at a higher or lower gain for the VPC-2100 compared to the CSRDF recognizer. The advantage of this wider dynamic range for input was that the VPC-2100 took less time to enroll than the CSRDF, for the same set of tokens.

APPENDIX G

Appendix G contains the responses to this report of those vendor(s) who chose to have them included in the non-proprietary section of the report. Their responses are reproduced here exactly as received.

Proprietary responses from vendors are not included in this report. Distribution of that information is limited to U.S. Government employees with a need to know for purposes of Government procurement. Such persons should contact Chief, Crew Station R&D Branch, M/S 243-4, Moffett Field, CA 94035.

ITT Corporation

ITT Defense Communications Division

Defense Technology Corporation

10060 Carroll Canyon Road
San Diego, California 92131
(619) 578-3080

Evaluation of Speech Recognizers For Use in Advanced Combat Helicopter Crew Station Research and Development

Response to Final Report ITT Defense Communications Division

December 1989

INTRODUCTION

ITT Defense Communications Division (ITTDCD) was pleased to have been selected for participation in Psycho-Linguistic Research Associates (PLRAs) evaluation of speech recognizers for possible upgrade of the U.S. Army Crew Station Research and Development Facility (CSRDF). We have been a leading player in DoD-sponsored speech recognition research and development programs since the late 1970's and have a laboratory with more than 30 full-time employees addressing the challenge of advancing the state-of-the-art of this technology. Our experience in developing and applying real-time speech recognition devices to tactical environments includes very successful participation in such flight tests as: the AFTI/F-16 Phase I program and two Concept Evaluation Programs on JOH-58C helicopters conducted by the U.S. Army Aviation Board, Ft. Rucker in 1987 and 1989 respectively.

Unfortunately, after reviewing the final draft of the PLRA report and after follow-up conversations with the author, ITTDCD has concluded that the results for the VRS 1280 are not representative of the typical performance end-users can obtain from this recognizer. The key reasons for this contention are summarized below.

1. Only two of the four speakers created voice templates for the isolated word tests using the full procedures recommended by ITTDCD. Moreover, none of the speakers created voice templates for the connected word test using our recommended procedures. This situtation arose because there was an incompatibility between the PD-100 template enrollment data base and the data which the VRS 1280 algorithm is designed to process to create the "best" possible set of templates for a given application. (A compromise between PLRA and ITTDCD permitted the collection of results for the two speakers who did conform in the isolated word tests.) As discussed later in this commentary, the test results, on average, show degradation in cases where the recommended procedures could not be used.

2. The single recognition syntax loaded to the VRS 1280 for all phases of testing explicitly modelled an acoustic environment in which only utterances with roughly 0.5 seconds or greater pause between words could be reliably recognized. As a consequence, the VRS 1280 could only reliably recognize the first word in most of the connected word strings which were presented. Despite this "handicap", the VRS 1280 scored significantly better than the CSRDF recognizer on the connected word tests. A simple change to the single recognition syntax we used for all phases of testing would have permitted both isolated word and connected word utterances to be reliably modelled. We are quite confident the performance for connected words would have been considerably higher, with little, if any, consequence to the isolated word results.

3. Results based upon an a-posteriori "optimized" rejection threshold were never obtained for the VRS 1280 because of a mutual misunderstanding between PLRA regarding the meaning of the scoring information which our recognizer produces and ITTDCD regarding PLRA's request for goodness of fit scores. As of this writing, the author has not provided ITTDCD with a debriefing to provide diagnostic information (see main body of Report - Phase III Data Analysis). Consequently, it is impossible for us to determine if performance would improve with an "optimized" threshold - but certainly the potential for improvement exists given the known difficulties of a-priori threshold optimization.

4. The PD-100 test was conducted in the fall of 1988. The VRS 1280 algorithm technology used in the test is no longer offered by ITTDCD. Several improvements have been made to more reliably detect and reject false alarms in noisy environments and dynamically adjust for speaking level changes between the templates and the test utterances. We are confident this new version of the firmware would greatly alleviate the false alarm problems mentioned in the report.

As a final introductory comment, the misunderstandings described above in points 2 and 3 were discussed with the author in late August 1989, after ITTDCD had reviewed the final draft. PLRA suggested two possible remedies. First, they agreed to retest the VRS 1280 (with the same technology used originally), if the sponsor would give PLRA authorization. In mid-September, we were informed that permission was not given because it could set a precedent for having to retest one or more of the other recognizers in response to vendor requests. Second, PLRA agreed to retest the VRS 1280 if ITTDCD funded the effort, with the understanding however that the sponsor in this case would not permit PLRA to modify the report to incorporate these additional results. PLRA did, of course, offer to make them available to ITTDCD for our own use. We did not consider these terms to be sufficiently attractive and are disappointed that an opportunity to include more representative results in the final report has been lost.

BACKGROUND

After reviewing the documentation provided by PLRA accompanying the invitation-to-participate, ITTDCD concluded that the PD-100 vocabulary and test measures collectively posed a significant challenge. Accordingly, we wanted to use our most advanced technology in the test. At the time we were asked to participate, this technology had only been implemented in the tactical recognizer we developed for the aforementioned Army-sponsored flight tests. New techniques had been incorporated to minimize false alarm errors induced by helicopter background noise, breath noise, etc. and to reduce sensitivity to changes in speaking level. Unfortunately, all of the recognizers in question were being utilized by ITTDCD and the U.S. Army Aviation Research and Development Activity (AVRADA) in preparation for these helicopter flight tests. Furthermore, PLRA required an IBM-PC compatible board for the test in order to conform to interface requirements with existing CSRDF speech recognition system hardware and software.

Rather than decline the invitation to participate in the tests and in view of PLRA's requirement, ITTDCD chose to use another of our single board recognizers, the VRS 1280/PC board, even though it had not been upgraded with these most recent advances made in our technology. As stated in the introduction, all of our single board recognizer standard products were subsequently enhanced with these upgrades.

TEMPLATE ENROLLMENT COMPROMISE

The template enrollment procedure for the PD-100 Evaluation was not compatible with the template generation algorithm used by VRS 1280/PC recognizer.

The recommended procedure for enrollment is described in Appendix C. This process begins with a set of speaker-pooled templates for the digits (0 - 9), which are provided with the VRS 1280/PC application development software. Three preliminary steps, requiring speech input, are then performed. These steps essentially represent a fully automatic methodology for "marking" a seed template for each word in the application vocabulary. A seed is needed because ITTDCD uses a template-based as opposed to inherently inaccurate heuristic endpoint detection technique to determine word boundaries as the application vocabulary is enrolled and word templates are created. Once a seed template for each vocabulary word exists, the final step in the enrollment process requires that each of the vocabulary words be spoken several times. Usually two or three repetitions of each word are sufficient to create a good averaged template to represent each vocabulary word. For continuous speech recognition, these final repetitions should be spoken in short continuous phrases which are representative of the application grammar in order to incorporate the co-articulation effects into each template.

A compromise, which is discussed in Appendix C, was reached concerning the enrollment process so that the VRS 1280/PC recognizer could be included in PLRA's test. The compromise involved recording the speech input data for the first three steps of the enrollment process for one male (M1) and one female (F2). The other two speakers, M3 and F4, skipped the first three steps and used the seed templates which were generated for M1 and F2.

The two speakers who did perform the full four step enrollment procedure obtained better performance than the two who only performed step four. Male speaker M1 performed better than the bootstrapped speaker M2 in each of the three cases - isolated word/quiet, isolated word/noise, and connected word. The same trend holds true for female speaker F2 compared with bootstrapped speaker F4 except for the case of isolated word/noise. Overall, the bootstrapped templates performed 7.5% (AOA) and 11.25% (OA) worse on average.

CONNECTED WORD TEST

One aspect of the PD-100 test was to study recognizer performance when presented with test tokens involving a series of connected words. The vocabulary templates used for the isolated word tests were also used for the connected word tests, as required by PLRA for all recognizers. In general, better performance could be obtained with template training data in which the speaker provided at least some connected speech involving the vocabulary items, to capture coarticulation effects. This is the recommended procedure for the VRS 1280. While this does modestly increase the template training time compared with a strictly isolated word training session, the positive impact to performance typically results in an acceptable tradeoff.

No attempt was made to configure the recognizer for both isolated and connected word recognition. A syntax was designed to explicitly model test tokens presented in isolation (i.e. with periods of background noise before and after all tokens). This syntax was set up such that roughly a 0.5 second gap between words was required for proper recognition processing. PLRA used this same syntax in the connected word tests.

It is not surprising that the human-normalized VRS 1280 Recognition Accuracy (RA) for connected words in this test was not representative of typical VRS 1280 performance. In general, the recognizer would be able to correctly recognize the first word in the string, but would miss most of the subsequent words because the fixed amount of time between words in the syntax model greatly exceeded the typical pause between words in the connected word test utterance data base.

Excellent Rejection Accuracy (JA) permitted the VRS 1280 to achieve a human-normalized Overall Accuracy (OA) and Adjusted Overall Accuracy (AOA) which still represented significantly better results than the CSRDF recognizer.

However, ITTDCD is convinced that if a syntax modelling connected word speech had been used on this portion of the PD-100 test, results far in excess of the mean AOA score for the four new recognizers tested would have been achieved. This syntax would have contained a model which would have permitted pauses of arbitrary length between words. Therefore, it would have also been effective on the isolated word components of the PD-100 test.

REJECTION THRESHOLD OPTIMIZATION

During PD-100 testing, PLRA wanted to collect a "goodness of fit" score, if available, for each recognized word. This score is presumably an indication of how close the test token matched the reported template. One commonly used approach involves applying a rejection threshold to this score in order to determine whether the reported template should be accepted as the response to a legal token, or rejected as an illegal token.

A second commonly used approach toward rejection is to examine the relative score of the best fitting template with the score of the second best fitting template. A large difference in these two scores may indicate a high probability that the word recognized is indeed correct; whereas, if the two scores are very close, the best-matched word is typically rejected. A threshold can also be applied to this difference score to control rejection rates.

The rejection algorithm used by the VRS 1280 recognizer is not based on relative scores between the first and runner up templates in the active vocabulary, but instead is based on relative scores between the best match in the active vocabulary and the best match in a set of "rejection filler" templates. In the PD-100 test, output was provided on the name of the best matched template from the active vocabulary, its score, and the score of the best rejection template. In addition, an indication was output if the best matched active vocabulary template was rejected.

Because of the difference between ITTDCD's rejection algorithm and the two aforementioned commonly used approaches, our technical representative told the author that more accurate rejection would be achieved by the comparison of the two scores provided rather than the application of a hard threshold to the score of the best matching active vocabulary template. This statement led the author to conclude that goodness of fit scores were not available from the VRS 1280.

PLRA calculated the performance percentages in two different ways. First, calculations were performed using the a-priori thresholds which were selected by the vendors. Later, for two recognizers, these calculations were performed again with different rejection thresholds based on the goodness of fit scores collected during the test. These "optimized thresholds" provided as good or better results than were achieved with the vendor selected thresholds.

After reading the final draft of this report, ITTDCD spoke with the author regarding why the "optimized threshold" analysis was not performed on the VRS 1280. We learned, at this time, of the mutual misunderstanding described above regarding the significance of the two scores our recognizer provided. The algorithm/software which PLRA used to compute the "optimized thresholds" could have been used for the VRS 1280 if ITTDCD had provided a simple program to reformat results to insure that the lower of the two scores appeared first. Thus, although not reflected in the results section, the VRS 1280/PC has the same potential for enhanced rejection accuracy using an optimized threshold.

It should also be pointed out that the technique described above is clearly not the only effective means of rejection. Other methods exist for the VRS 1280 which were not utilized in this test because they actually suppress output of the best matching active vocabulary template upon rejection of an illegal test token.

APPENDIX C - PREPARATION AND ENROLLMENT TIMES

The time required for preparation and enrollment reported in Appendix C greatly exceed the typical preparation and enrollment times for standard speech recognition applications using the VRS 1280 recognition system.

Much of the time spent in preparation and set up involved unsuccessful attempts to enroll the numbers vocabulary. The attempts were not successful because of the deviation from ITTDCD's standard enrollment procedures. This accounted for approximately half the reported total preparation time.

Time spent on preparation of data files needed for enrollment of both the numbers and PD-100 vocabulary were greater than that typically required. As described in Appendix C, 17 files were modified or created for each of the PD-100 vocabularies in order to accommodate the taped speech data. Most applications, including the two live helicopter flight tests in which ITTDCD has participated, require only six files to be created. In addition, much detail was provided to PLRA personnel about the creation and rationale behind each preparation step. The times reported in this report are more typical of both tutorial and application preparation times combined. The author confirmed in a personal communication that part of the assessment involved the measurement of learning and set-up time for an individual knowledgeable about speech recognition but unfamiliar with the particular recognition device and development software.

Enrollment of the numbers vocabulary was successful after incorporating our recommended procedures. This included preparing some data files, recording the numbers vocabulary within carrier phrases, the actual enrollment and verification of these words, and finally the testing of these numbers in vocabulary subsets similar to the PD-100 vocabulary. Again, much of this time is not required in a standard application and was performed only for PLRA data collection.

Enrollment of the PD-100 vocabulary was also very time consuming. This was due to a variety of reasons. One reason was the fact that all speech used for this test was produced from a recorded database. Much time was spent advancing or rewinding the tape to the proper alignment for enrollment token presentation, whereas, during live enrollment, a speaker can produce speech immediately upon being prompted. This was especially true when templates had to be enrolled more than once (e.g. it was sometimes difficult to find a particular enrollment token). As was the case in the numbers vocabulary testing, much time was spent collecting data which is not required in ITTDCD's standard procedures, but only performed at PLRA's request.

To cite a specific practical example, during the aforementioned helicopter flight tests which were conducted at Ft. Rucker in March of 1989, the six pilots who participated in the tests averaged less than two hours each to generate templates for over 200 vocabulary words using the recommended four step enrollment procedure.

APPENDIX C - PROBLEMS

With reference to the section entitled "Problems" of Appendix C, the following comments are appropriate:

1. ITTDCD has offered an MS/DOS compatible version of the VRS 1280/PC system since June 1989.

2. In June 1989, ITTDCD also released a second-generation version of our template training software. The version used by PLRA during the PD-100 test has been discontinued. The new version addresses several of the well-taken criticisms cited by the PLRA report. Visual feedback is now provided on the audio signal input level during template generation. A sophisticated scoring algorithm for judging template quality has been incorporated which alleviates most of the decision making from the speaker. The man-machine interface has been significantly changed and is now much more novice-user friendly. Feedback from customers who have upgraded to this second-generation version has been uniformly positive regarding both ease-of-use and performance.

3. As stated earlier, the number of files which ITTDCD's technical representative created for vocabulary enrollment and testing was much higher than typical for the version of software used in the PD-100 test. A new, more user-friendly application development tool package for creating the six files typically needed was also introduced in 1989.

Marconi
Speech & Information Systems

PORTSMOUTH

September 1989

THE EVALUATION OF SPEECH RECOGNISERS FOR USE IN

ADVANCED COMBAT HELICOPTER CREW STATION RESEARCH AND DEVELOPMENT

COMMENTS ON THE FINAL DRAFT ON THE MARCONI MACROSPEAK

We have read the report on the design, procedures and results of these tests with great interest. On the basis of the information provided, we consider that it is a fair and comprehensive test of Macrospeak and the other recognisers.

In the presentation of the results with Macrospeak, the addition of the following information would have made the report even more useful:

1. A statement about whether the results were absolute or human normalised.

2. A comment on the exceptionally poor performance of speaker M3 relative to the other three speakers in both the single-word tests and to his own performance in the connected-word test which is generally more difficult.

3. A statement about whether the Macrospeak results were obtained using the threshold set at the vendor's recommended value or the optimised value. The improvement in the average performance of the recognisers provided by the optimised threshold was particularly interesting. We would be glad to know what the individual improvement for Macrospeak was.

Comments of the Review Table 1

Table 1 in the Report is drawn from the responses to the Phase II questionnaire concerning recogniser response time for both isolated words and triple word phrases.

We were concerned to provide absolutely honest responses to the questionnaire and we found the questions in this section to be open to various interpretations. In particular, we find it difficult to know what is meant by the "end" of a word. Our definition may well be different from that used by PLRA or by the other vendors. The slow decay in energy at the ends of some words could lead to differences of up to 300 ms in judgments of word endings. Under noisy conditions, the uncertainty may be even greater .

In most conditions, our users find the response times of Macrospeak to be adequate. We suggest that the only fair way to compare the performances of the various recognisers in this respect would be to measure response times on identical material.

Brian M. Nicholas, Speech Group Manager
Melvyn J. Hunt, D Phil, Chief of Speech Research
Ian Galletti, Marketing Manager
Martin G. Abbott, Engineering Manager

NASA
National Aeronautics and Space Administration

Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-177547 USAAVSCOM TM-90-A-001		

4. Title and Subtitle	5. Report Date
Evaluation of Speech Recognizers for Use in Advanced Combat Helicopter Crew Station Research and Development	March 1990
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Carol A. Simpson	A-90090
	10. Work Unit No. 505-61-51
9. Performing Organization Name and Address	11. Contract or Grant No.
Psycho-Linguistic Research Associates, 485 Summit Springs Rd., Woodside, CA 94062 and U. S. Army Aeroflightdynamics Directorate, Ames Research Center, Moffett Field, CA 94035-1000	NAS2-12425
	13. Type of Report and Period Covered: Contractor Report
12. Sponsoring Agency Name and Address	14. Sponsoring Agency Code
National Aeronautics and Space Administration Washington, DC 20546-0001 and U. S. Army Aviation Systems Command, St. Louis, MO 63120-1798	

15. Supplementary Notes

Point of Contact: Nancy Bucher, Ames Research Center, MS 243-4, Moffett Field, CA 94035-1000
(415) 604-5161 or FTS 464-5161

16. Abstract

The U. S. Army Crew Station Research and Development Facility uses vintage 1984 speech recognizers. An Evaluation was performed of newer off-the-shelf speech recognition devices to determine whether newer technology performance and capabilities are substantially better than that of the Army's current speech recognizers. The Phonetic Discrimination (PD-100) Test was used to compare recognizer performance in two ambient noise conditions: quiet office and helicopter noise. Test tokens were spoken by males and females and in isolated-word and connected-word mode. Better overall recognition accuracy was obtained from the newer recognizers. The report lists recognizer capabilities needed to support the development of human factors design requirements for speech command systems in advanced combat helicopters.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Speech recognizers, Crew station research and development, Helicopters, Performance evaluation, Speech input, Phonetic discrimination	Unclassified-Unlimited Subject Category - 05

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	91	A05

NASA FORM 1626 OCT 86

For sale by the National Technical Information Service, Springfield, Virginia 22161