[redacted]

N93-72620

58-32
176344

AUTOMATIC SPEECH RECOGNITION RESEARCH AT NASA-AMES RESEARCH CENTER

Clayton R. Coler
NASA-Ames Research Center
Moffet Field, CA 94035

Robert P. Plummer
University of Utah
Salt Lake City, UT 84112

Edward M. Huff
NASA-Ames Research Center
Moffet Field, CA 94035

Myron H. Hitchcock
Computer Sciences Corporation
Mountain View, CA 94043

PRECEDING PAGE BLANK NOT FILMED

Automatic speech recognition (ASR) is being investigated at NASA-Ames Research Center as part of a broad program in Flight Management Systems research. The goal of the Flight Management program is to develop a base of practical knowledge and experience concerning pilot information and control requirements for future highly automated commercial flight systems [1]. The motivation for this research is concern that the air traffic environment is becoming highly congested and will eventually become saturated unless some means is found to improve the overall precision and scheduling of flights, particularly in the dense terminal area.

Various display and control devices are being investigated as aids to the crew of future aircraft. Potential information displays include multifunction, area-navigation, moving map, collision warning, traffic situation, ground proximity, system status, and various attitude displays. Virtually all of these displays involve selection features such as orientation, scaling, symbology, or numerical parameter options that may be left to the pilot to determine. Option selection will be determined in part by the ease or difficulty that the pilot will have in making his choices, and the degree to which these activities may interfere with more important pilot functions. Speech technology offers opportunities for increasing the effectiveness of pilot-system interaction, and both speech recognition and speech synthesis are being considered: the former as a potential input alternative to numerous and complex keyboard arrangements, and the latter as an optional output medium for presenting flight-critical information. This paper describes speech recognition work at NASA-Ames Research Center. Near-term ASR testing and evaluation in a motion simulator, and subsequent

PRECEDING PAGE BLANK NOT FILMED

PAGE 142 INTENTIONALLY BLANK

Page Intentionally Left Blank

(ties are resolved arbitrarily). Assuming that the vocabulary words are equally likely to occur and that all misclassifications are equally costly, the maximum likelihood decision is obtained by using as the discriminant functions

$$g_i(X) = p(X|i), \quad i = 1, \ldots\ldots, N.$$

That is, the ith discriminant function is the likelihood of pattern, X, with respect to category, i. Since the patterns are 120-bit binary vectors, and assuming the pattern components to be statistically independent, then

$$g_i(X) = \prod_{j=1}^{120} p(x_j|i), \ i = 1, \ldots, N, \text{ where}$$

$X = (x_1, \ldots\ldots, x_{120})$ and each x_j is 0 or 1.

For computational purposes, the discriminant functions actually used are

$$g_i(X) = \log \prod_{j=1}^{120} p(x_j|i) = \sum_{j=1}^{120} \log p(x_j|i).$$

The logarithms, which are computer by table look-up, do not affect the classification since the log function is monotonic-increasing.

It remains to state how the probabilities, $p(x_j|i)$ are found. This is done in advanced by collecting training samples, $Y^{i,1}, \ldots Y^{i,m}$ for each vocabulary word, i, and using the following as estimates for the probabilities:

$$p(x_j = 1|i) = \frac{\sum_{k=1}^{M} y_j^{i,k}}{M}$$

and $p(x_j = 0|i) = 1 - p(x_j = 1|i)$.

The number of training samples, M, is usually between 10 and 25.

Three additional features have been added to the maximum likelihood algorithm. The first derives from the fact that even practiced speakers change some of their pronunciations slightly over time. Compensation from these changes may be accomplished by updating the discriminant function probabilities. In situations where the algorithm receives feedback stating the correctness of its classifications, then if pattern, Z, is correctly classified as belonging to category, m, the probabilities $p(x = 1|m)$ are replaced by

$$a \cdot p(x_j = 1|m) + (1-a) \cdot z_j \ , \text{ for } 0<a<1 \text{ and } j=1, \ldots, 120.$$

PRECEDING PAGE BLANK NOT FILMED

PAGE 144 INTENTIONALLY BLANK

This exponential smoothing technique can be used to make the algorithm more or less responsive to changes in pronunciation by varying the value of the variable a. A typical value is 0.94.

A second feature allows the algorithm to "reject" utterances for which classification is uncertain, rather than risk misclassification. The measure of uncertainty is a simple one: the ratio of the second highest score to the highest. The nearer the ratio to 1.0, the more uncertain the classification. The threshold at which rejection takes place is a parameter of the algorithm.

Finally, in many flight systems applications, the command language spoken by the user of the speech recognition system has a simple (finite-state) syntax. For example, after the command "landing gear," the only meaningful utterances might be "up," "down," and "status." By doing recognition only against the subset of vocabulary words that is valid at each point in a command string, both the accuracy and efficiency of the algorithm are increased. A tree-like structure is used to associate with each vocabulary word that set of words that can follow it in a command string. As each word is recognized, the result is used to guide the traversal of the tree.

The current version of the Ames speech recognition system runs in real time on a PDP 11/10 computer. Encoding an utterance into 120-bit form requires about 0.3 seconds, and recognition requires an additional 0.2 seconds for small vocabularies or subsets.

RECOGNITION ALGORITHM EVALUATION

A standard set of data was needed for evaluation of the various recognition algorithms. To meet this requirement, 20 untrained male speakers used the 10-word digit vocabulary to provide a total of 1250 utterances for each speaker. The data for each speaker were collected in 5 blocks of 250 utterances each during a single 1 hour 30 minute session.

During data collection, the 120-bit pattern derived from each utterance was transmitted directly from the VCS to a PDP-12 computer for storage on magnetic tape. All data were later transferred to a Xerox Sigma 9 computer where the various recognition algorithms were used to process the data. During data processing, the first block of data for each speaker was used to train the recognition software and then recognition was attempted on the remaining 4 blocks of data for that speaker; thus 1,000 recognition utterances were processed for each speaker. The same data were repeatedly processed to provide independent evaluation of each recognition algorithm. The results are shown in Table I.

Pg. 144 ✓

PRECEDING PAGE BLANK NOT FILMED

TABLE 1

SPEECH RECOGNITION ACCURACY FOR UNTRAINED GROUP: 10-WORD (DIGIT) VOCABULARY

(20 SPEAKERS; 1000 UTTERANCES EACH SPEAKER)

RECOGNITION ALGORITHM	PERCENTAGE CORRECT		
	MEAN	STANDARD DEVIATION	RANGE
$VCS_{(5)}$	87.6	6.4	64.5 — 94.4
$SW_{(5)}$	95.0	2.0	91.1 — 98.6
$SW_{(10)}$	96.9	1.7	93.4 — 98.9
$SUBMAX_{(10)}$	99.4	0.3	98.8 — 99.8
$SUBMAX_{(25)}$	99.5	0.3	98.7 — 99.9
$SUBMAX_{(25)}$ REJ*	99.9	0.1	99.6 — 100.0
*PERCENTAGE REJECTED:	5.0	2.3	1.7 — 9.4

The recognition algorithms in Table 1 are listed in order of increasing efficiency. The first algorithm, VCS(5) requires 5 training samples of each command, and simulates the hard-wired algorithm of the VCS (for comparison with initial VCS test results). This least effective recognition algorithm correctly classified 87.6% of 20,000 utterances. In contrast, 99.9% of 19,000 utterances (1000 utterances were rejected) were correctly classified by the most effective algorithm, designated SUBMAX(25)REJ (incorporating all the features discussed above and trained on 25 samples of each word). Thus a mean improvement of 12.3% correct was gained over the basic recognition technique. Syntax was not involved in any of the 10-word testing; all recognition was done against the entire vocabulary.

The greatest sequential improvement, 7.4% correct, was gained by use of the algorithm designated SW(5), a maximum likelihood algorithm without rejection or updating, and trained on 5 samples of each word. Doubling the size of the training data set with the same algorithm, SW(10), produced an additional gain of 1.9% correct. Use of rejection and updating with the same size training data set, algorithm SUBMAX(10), provided a further gain of 2.5% correct, and use of all available training data, algorithm SUBMAX(25), provided an additional gain of 0.1% correct. By using algorithm SUBMAX(25)REJ, which rejected the 5% of utterances with the greatest uncertainty of classification, an additional 0.4% correct was gained. (The value of the rejection threshold that produced 5% rejection was determined empirically.)

Both the standard deviation and the range of mean percentage correct (speaker with highest mean minus speaker with lowest mean) consistently decreased as increasingly efficient algorithms were used. The standard deviation decreased from a maximum of 6.4% to a minimum of 0.1% correct, and the range decreased from a maximum of 29.9% to a minimum of 0.4% correct. Large relative reductions in standard deviation (from 6.4% to 2.0% correct) and range (29.9% to 7.5% correct) resulted from use of the SW95) algorithm. The largest relative reductions in standard deviation (from 1.7% to 0.3% correct) and range (from 5.5% to 1.0% correct) resulted from use of the SUBMAX(10) algorithm (with dynamic updating).

The principal achievement of the Ames recognition algorithm development work is that recognition accuracy for the least successful of the 20 speakers was improved to 98.7% correct without rejecting any of his utterances. For successful flight applications, a recognition system must perform well for even the least proficient speaker.

Since asking a pilot to repeat a command seems preferable to misclassification [3], the ability to reject in cases of uncertainty is desirable. With 5% rejection, recognition accuracy was further improved to 99.6% correct for the least proficient speaker.

After achieving a high level of recognition accuracy for the 10-word digit vocabulary, it was desirable to evaluate recognition accuracy of the system with a larger vocabulary suitable for flight system use.

RECOGNITION ALGORITHM ACCURACY ON A 100-COMMAND VOCABULARY

Although a vocabulary larger than 10 commands would be required for most flight systems applications, only a few of the commands may be needed at any given time during a mission. The commands may be assigned to different vocabulary subsets according to their sequential use in the mission. The pilot would have access to any command at all times simply by executing the proper access sequence for that command. For example, a 58-command vocabulary selected for use in a fixed-base flight simulator mission consists of 17 different vocabulary subsets; the smallest subset contains 3 commands, and the largest contains 12. A syntax structure was imposed to develop branching chains of command by sequentially combining appropriate commands from various subsets to yield a total of more than 46,000 unique and potentially meaningful sequences (from the original 58-command vocabulary). In addition to increasing the number of unique command possibilities, the syntax structuring method also reduces the number of active commands in the recognition set (min. = 3; max. = 12) at any given point in time, thus reducing the complexity of the recognition problem and increasing the probability of maintaining a high level of recognition accuracy over the entire mission simulation [4].

To evaluate recognition accuracy on a large vocabulary, a 100-command flight vocabulary constructed for use in a full mission (take-off-to-landing) simulation was selected. The 100-command vocabulary is shown in Table 2.

A group of ten untrained male speakers each used the 100-command vocabulary to provide 25 training utterances (used to train the recognition algorithms) and 100 recognition utterances of each command, a group total of 100,000 recognition utterances.

Overall Processing Results

Recognition results for overall processing without the use of syntax are shown in Table 3. Recognition accuracy for the entire vocabulary was 93.2% correct without rejection and 95.7% correct when 5% of the utterances were rejected. Results for the 10 digits within the overall processing are shown for comparison with previous 10-digit vocabulary results.

The command language syntax for this vocabulary groups the commands into 15 subsets ranging in size from 3 to 10 commands as shown in Table 4.

TABLE 2

100-COMMAND FLIGHT SIMULATION VOCABULARY

1. AIRPORTS
2. ARRIVAL TIME
3. AIRSPEED
4. ALTITUDE
5. ALPHA
6. AUTO
7. AUTOPILOT
8. BETA
9. BLANK
10. BY
11. CAPTURE
12. CENTER
13. CHANGE
14. CHANNEL
15. CHART
16. CHECK LIST
17. CLEAR
18. CLIMB
19. COMMUNICATION
20. COUPLE
21. COURSE
22. DELTA
23. DEGREES
24. DESCEND
25. DISPLAY
26. DOWN
27. EAST
28. EIGHT
29. EMERGENCY
30. ENGAGE
31. ENTER
32. ERROR
33. EXECUTE
34. FIVE
35. FLIGHT PLAN
36. FLY TO
37. FOUR
38. FREQUENCY
39. GAMMA
40. GLIDESLOPE
41. GO
42. GO AROUND
43. HEADING
44. HOLD
45. HORIZONTAL
46. HUNDRED
47. I.D.
48. IN
49. INDEX
50. INITIALIZE
51. INPUT
52. LARGER
53. LEFT
54. MAP
55. MASTER
56 M.L.S.
57. MERGE
58. MINUS
59. NAV AUTO
60. NAVIGATION
61. NEGATIVE
62. NINE
63. NORTH
64. OFF
65. ON
66. ONE
67. OUT
68. OVER
69. PLUS
70. POINT
71. POSITIVE
72 PREDICTOR
73. REFERENCE
74. RIGHT
75. SCALE
76. SECTOR
77. SELECT
78. SET
79. SEVEN
80. SIX
81. SMALLER
82. SOUTH
83. SPEED
84. TERRAIN
85. THOUSAND
86. THREE
87. TIME
88. TRACK
89. TURN
90. TWO
91. UNDER
92. UP
93. VECTOR
94. VELOCITY
95. VERIFY
96. VERTICAL
97. V.O.R.
98. WAYPOINT
99. WEST
100. ZERO

TABLE 3

SPEECH RECOGNITION ACCURACY FOR 10 UNTRAINED SPEAKERS: 100-WORD FLIGHT SIMULATION VOCABULARY

RECOGNITION RESULTS FOR OVERALL PROCESSING

RECOGNITION ALGORITHM	PERCENTAGE CORRECT FOR ENTIRE VOCABULARY		PERCENTAGE CORRECT FOR DIGIT SUBSET	
	MEAN	RANGE	MEAN	RANGE
SUBMAX$_{(25)}$	93.2	89.3 – 96.8	91.8	88.4 – 93.7
SUBMAX$_{(25)}$ REJ*	95.7	93.2 – 98.1	94.7	92.0 – 96.4
*FREQUENCY OF REJECTION	5.0	2.8 – 7.1	6.0	3.2 – 8.2

TABLE 4

FREQUENCY OF SUBSET SIZES:
100-COMMAND FLIGHT SIMULATION VOCABULARY

SUBSET SIZE	FREQUENCY
3 WORDS	2
4 WORDS	3
5 WORDS	1
6 WORDS	3
7 WORDS	—
8 WORDS	—
9 WORDS	1
10 WORDS	5
	TOTAL = 15 SUBSETS

Subset Processing Results

Recognition results for subset processing are shown in Table 5. For the entire vocabulary, subset recognition accuracy without rejection is 98.6% correct, and is higher than the corresponsing overall processing result by 5.4% correct, while the range has decreased from 7.5% to 1.8% correct. With 5% rejection, subset recognition accuracy is increased by 1.0% correct to 99.6% correct, and the range is decreased to 0.5% correct. Results for the 10 digits within the subset processing are shown for comparison with previous 10-digit results.

Tables 6, 7, and 8 show the commands, subset processing results without rejection, and rank order of recognition accuracy for each of the 15 subsets. The subset processing results shown in Tables 6, 7, and 8 are summarized in Table 9, together with corresponding results from the overall processing.

During overall processing, recognition was done against the entire 100-command vocabulary. (During subset processing, the recognition decision for each utterance was based on comparison to only 3, 4, 5, 6, 9, or 10 commands, depending upon the size of the active subset.) The overall processing results shown in Table 9 were then obtained by simply combining and averaging individual command results by subset group. (The subset groups are shown in Tables 6, 7, and 8.)

The rank order results for subset processing shown in Table 9 are in general agreement with the expectation that recognition accuracy will decrease as subset size is increased. When recognition accuracy for a subset of small size was considerably below the levels obtained with larger subsets, the disparity was usually attributable to frequent misclassification between two commands in the small subset. For example, although Subset L contains only 4 commands, the resulting mean percentage correct by subsetting was lower than corresponding results for eight subsets of larger size. The primary reason for error within Subset L was misclassification between the commands "alpha" and "delta" (see Table 7).

Digit Subset Results

The 10-digit subset, subset D, ranged 14th of the 15 subsets in recognition accuracy by subset processing and 13th by overall processing. Of the five 10-command subsets, subset D recognition accuracy ranked 4th by both processing methods. These results demonstrate that the 10 digits comprise a relatively difficult 10-word ASR vocabulary. Since accurate recognition of the 10 digits is essential for most potential flight systems applications, the digits appear to be an excellent small vocabulary for use in recognition accuracy evaluation of an ASR system proposed for flight systems use.

TABLE 5

SPEECH RECOGNITION ACCURACY FOR 10 UNTRAINED SPEAKERS: 100-WORD FLIGHT SIMULATION VOCABULARY

RECOGNITION RESULTS FOR SUBSET PROCESSING

RECOGNITION ALGORITHM	PERCENTAGE CORRECT FOR SUBSETS: ENTIRE VOCABULARY		PERCENTAGE CORRECT FOR DIGIT SUBSET	
	MEAN	RANGE	MEAN	RANGE
SUBMAX$_{(25)}$	98.6	97.5 – 99.3	98.2	97.1 – 99.1
SUBMAX$_{(25)}$ REJ*	99.6	99.3 – 99.8	99.6	98.9 – 99.9
*FREQUENCY OF REJECTION	5.0	2.7 – 7.3	7.4	4.1 – 14.0

TABLE 6

SUBSETS FOR 100-COMMAND FLIGHT SIMULATION VOCABULARY

MASTER PAGE (LEVEL 1)
SUBSET A (6 COMMANDS)

7. AUTOPILOT
15. CHART
16. CHECK LIST
19. COMMUNICATION
35. FLIGHT PLAN
60. NAVIGATION

98.7% CORRECT
RANK = 9

AUTOPILOT COMMANDS (L2)
SUBSET B (4)

4. ALTITUDE
43. HEADING
83. SPEED
96. VERTICAL

99.15% CORRECT
RANK = 3.5

NAVIGATION COMMANDS (L2)
SUBSET C (4)

2. ARRIVAL TIME
56. M.L.S.
59. NAV AUTO
97. V.O.R.

98.725% CORRECT
RANK = 7

COMMUNICATION AND
AUTOPILOT COMMANDS (L2)
SUBSET D (10)

28. EIGHT
34. FIVE
37. FOUR
62. NINE
66. ONE
79. SEVEN
80. SIX
86. THREE
90. TWO
100. ZERO

98.21% CORRECT
RANK = 14

CHART COMMANDS (L2)
SUBSET E (6)

1. AIRPORTS
54. MAP
72. PREDICTOR
75. SCALE
84. TERRAIN
94. VELOCITY

99.03% CORRECT
RANK = 5

TABLE 7

SUBSETS FOR 100-COMMAND FLIGHT SIMULATION VOCABULARY

AUTOPILOT: ALTITUDE, HEADING, AND SPEED COMMANDS (L3)
SUBSET F (9 COMMANDS)

44. HOLD
46. HUNDRED
51. INPUT
58. MINUS
69. PLUS
70. POINT
73. REFERENCE
77. SELECT
85. THOUSAND

98.48% CORRECT
RANK = 10

AUTOPILOT: VERTICAL COMMANDS (L3)
SUBSET G (6)

18. CLIMB
24. DESCEND
45. HORIZONTAL
48. IN
67. OUT
89. TURN

99.15% CORRECT
RANK = 3.5

NAVIGATION: M.L.S. COMMANDS (L3)
SUBSET H (5)

14. CHANNEL
30. ENGAGE
40. GLIDESLOPE
64. OFF
65. ON

98.86% CORRECT
RANK = 6

NAVIGATION: V.O.R. COMMANDS (L3)
SUBSET I (3)

21. COURSE
38. FREQUENCY
47. I.D.

99.53% CORRECT
RANK = 2

CHART: SCALE COMMANDS (L3)
SUBSET J (3)

52. LARGER
55. MASTER
81. SMALLER

99.7% CORRECT
RANK = 1

CHART: MAP COMMANDS (L3)
SUBSET K (10)

12. CENTER
26. DOWN
27. EAST
49. INDEX
53. LEFT
63. NORTH
74. RIGHT
82. SOUTH
92. UP
99. WEST

97.98% CORRECT
RANK = 15

COMMUNICATION COMMANDS (OPTIONAL-L2)
SUBSET L (4)

5. ALPHA
8. BETA
22. DELTA
39. GAMMA

98.4% CORRECT
RANK = 13

TABLE 8

SUBSETS FOR 100-COMMAND FLIGHT SIMULATION VOCABULARY

NAVIGATION: NAV AUTO COMMANDS (L3)
SUBSET M (10 COMMANDS)

10. BY
11. CAPTURE
23. DEGRESS
36. FLY TO
50. INITIALIZE
57. MERGE
76. SECTOR
87. TIME
88. TRACK
98. WAYPOINT

98.72% CORRECT
RANK = 8

NAVIGATION: ARRIVAL TIME COMMANDS (L3)
SUBSET N (10)

3. AIRSPEED
6. AUTO
20. COUPLE
25. DISPLAY
29. EMERGENCY
42. GO AROUND
68. OVER
78. SET
91. UNDER
93. VECTOR

98.43% CORRECT
RANK = 12

NAVIGATION: GENERAL COMMANDS (L4)
SUBSET O (10)

9. BLANK
13. CHANGE
17. CLEAR
31. ENTER
32. ERROR
33. EXECUTE
41. GO
61. NEGATIVE
71. POSITIVE
95. VERIFY

98.47% CORRECT
RANK = 11

TABLE 9

RANK ORDER OF SUBSETS: 100-COMMAND VOCABULARY

SUBMAX$_{(25)}$ – NO REJECTION

SUBSET	MEAN PERCENTAGE CORRECT BY SUBSETTING	RANK BY SUBSETTING	RANK OVERALL	MEAN PERCENTAGE CORRECT OVERALL
J (3)	99.70	1	2	95.56
I (3)	99.53	2	1	95.60
B (4)	99.15	3.5	5	94.48
G (6)	99.15	3.5	12	92.45
E (6)	99.03	5	8	93.67
H (5)	98.86	6	14	91.12
C (4)	98.725	7	3	95.53
M (10)	98.720	8	10	92.57
A (6)	98.70	9	4	94.75
F (9)	98.48	10	6	94.33
O (10)	98.47	11	9	93.56
N (10)	98.43	12	7	94.03
L (4)	98.40	13	11	92.55
D (10)	98.21	14	13	91.79
K (10)	97.98	15	15	91.02

Recognition accuracy and rank order results for the 10 digits are shown in Table 10. Recognition accuracy was highest for "six" and lowest for "five" by both subset and overall processing. Among individual commands in the entire 100-command vocabulary, "six" ranked 11.5 in recognition accuracy by subsetting and 2nd by overall processing. For comparison, "smaller" (Subset J) ranked 1st by subsetting (37th by overall) with 99.9% correct, and "waypoint" (Subset M) ranked 1st by overall processing (11.5) by subsetting) with 98.3% correct. In contrast, the digit "five" ranked 100th by both subsetting and overall processing. Thus differences in recognition accuracy within the digit subset cover nearly the entire range of results obtained over all 100 individual commands.

Within the digit subset, the most of the recognition misclassifications in subset processing occurred between "five" and "nine". Of the total misclassifications of "five" nearly 80% had been recognized as "nine", while more than 80% of all misclassifications of "nine" had been recognized as "five". For the digit subset, recognition accuracy could be improved by having speakers pronounce "nine" as "niner" (they did not do so in this study). Similar pronunciation changes could be made in other subsets where misclassification between two commands accounts for a high percentage of the total error. In some cases, a new command should be substituted for one of the problem commands and the evaluation process repeated.

The mean percentage difference in recognition accuracy between subset and overall processing for each digit is shown in Table 11. The relative cost of attempting to recognize a digit while protected within its own subset vs. leaving it unprotected in the overall processing is indicated by the value shown for each digit.

New vocabularies proposed for flight systems use may be evaluated by the process described for evaluation of the 100-command vocabulary. Thus any serious incompatibilities between commands within a subset may be identified and corrected, and recognition accuracy for the entire vocabulary can be determined prior to actual use in a flight system.

Results of the 100-command vocabulary evaluation indicate that recognition accuracy of the Ames speech recognition system is sufficiently high to permit operational testing.

OPERATIONAL TESTING IN NOISE AND VIBRATION

Since both noise and vibration are present in flight environments and since both threaten to reduce the high levels of recognition accuracy obtainable under laboratory conditions, a system recognition accuracy evaluation is scheduled for several different conditions of noise or vibration.

TABLE 10

RANK ORDER OF DIGITS: 100-COMMAND VOCABULARY

SUBMAX$_{(25)}$ – NO REJECTION

DIGIT	MEAN PERCENTAGE CORRECT BY SUBSETTING	RANK BY SUBSETTING	RANK OVERALL	MEAN PERCENTAGE CORRECT OVERALL
SIX	99.5	1	1	98.0
THREE	99.3	2	5	94.2
ZERO	99.8	3	2	96.7
FOUR	98.7	4.5	4	95.2
SEVEN	98.7	4.5	7	88.2
ONE	98.2	6.5	6	94.0
TWO	98.2	6.5	3	96.3
EIGHT	97.8	8	9	85.5
NINE	96.6	9	8	87.8
FIVE	96.3	10	10	81.9

TABLE 11

PERCENTAGE DIFFERENCES FOR DIGITS: 100-COMMAND VOCABULARY

SUBSET vs. OVERALL PROCESSING

DIGIT	MEAN PERCENTAGE DIFFERENCE
SIX	1.5
TWO	1.9
ZERO	2.1
FOUR	3.5
ONE	4.2
THREE	5.1
NINE	8.8
SEVEN	10.5
EIGHT	12.2
FIVE	14.4

To provide a more realistic test of the Ames ASR system than would be obtained if speakers simply voiced commands in a noise or vibrating environment, voice command data will be collected while the test subjects perform a continuous tracking task. Comparable keyboard entry data will also be collected.

Background

During flight a pilot must perform a variety of tasks, most of which may be assigned to one of two broad categories: (1) tracking, e.g., following a glideslope or making a turn, and (2) system interaction, e.g., setting an autopilot heading or changing a radio frequency. The use of digital computers on board aircraft (primarily for automatic subsystems control) is already a reality, and in future aircraft the role of the computer will be large [1,5]. Thus tasks that involve interacting with onboard avionics systems are increasingly becoming tasks of man-computer interaction. Effective utilization of the computer's information processing capabilities requires careful study and critical testing to provide design guidelines for the man-computer interface. In particular, a crew member must be able to provide inputs in a manner that is

(1) accurate,
(2) tolerant of errors and updates,
(3) rapid enough to meet the demands of the task at hand,
(4) natural and convenient, so that use of the input system does not add significantly to the user's workload, and
(5) interruptable.

Since speech has the potential for equaling or exceeding conventional keyboard input capabilities in meeting these requirements [6], the operational testing is designed to assess the relative effectiveness of voice and keyboard input systems in stationary, noisy, and vibrating environments.

Plan of Study

Pilots will act as test subjects and will perform a single-axis compensatory tracking task while being exposed to various levels of noise. While tracking, the pilots will concurrently make discrete numerical data entries, upon command, using one of two input media: voice or keyboard. Following completion of the required number of noise evaluation test sessions, the pilots will perform the same tasks for an equivalent number of test sessions under various levels of vibration. The pilots will be selected from the general aviation, military, and airline pilot populations.

All noise and vibration test data will be collected in the Ames Vertical Acceleration and Roll Device (VARD). The noise and vibration will simulate conditions occurring on board aircraft. Four noise conditions will be tested:

(1) No Noise,
(2) Helicopter noise at 90 dB,
(3) Helicopter noise at 100 dB,
(4) Random ("white") noise at 100 dB.

For each pilot, no vibration testing will be imposed until all noise testing has been completed. Four vibration conditions will then be simulated and tested:

(1) No vibration,
(2) Smooth jet transport cruise,
(3) Rough jet transport cruise,
(4) Helicopter cruise.

For all test conditions, the primary performance measures will be tracking error and the accuracy and speed of voice and keyboard data entries. Assessment of the relative effect on tracking performance of making voice vs. keyboard entries under the various noise and vibration conditions is of particular interest.

Depending upon initial noise and vibration test results, additional evaluation may be desirable for other noise or vibration conditions, or combinations of noise and vibration. When favorable ASR system results have been achieved under the noise and vibration conditions expected in flight, the system will be ready for flight testing.

FLIGHT APPLICATIONS

The hardware requirements for a flyable ASR system are quite different from the requirements for a laboratory system. Selection of a hardware system that will be fully operative under a variety of flight conditions is essential for later success in flight.

The Ames ASR Flight System

The primary component of this Ames ASR system is a Rolm 1603A ruggedized military computer. The entire system is contained in a 7.6" x 10.1" x 20.4" chassis and weighs about 60 lbs. The system meets military physical operating standards set by MIL-E-5400 Class II and MIL-E-16400 Class I specifications with respect to operating temperature, vibration, shock, humidity, altitude, and radio frequency interference. It contains 32,764 16-bit words of memory, central processor with extended arithmetic unit, several general purpose input/output ports, and a microprocessor-based speech input board. This board accepts microphone inputs and performs shaping, spectral analysis, word boundary detection, nonlinear time normalization, and coding completely independent from the 1603A processor. The 1603A accepts 128-bit patterns from the speech input board, performs training and recognition functions, and directs overall control of a flight experiment. For a single user, structured vocabularies of up to 200 words may be accommodated.

Potential Flight Applications

Although the potential flight applications of ASR systems are numerous, those applications that seem particularly desirable have general characteristics that would allow the ASR system to (1) reduce the difficulty (and risk) of performing high-wordload manual control procedures during critical phases of flight, (2) minimize eye-hand coordination problems that often accompany a heavy manual control burden, and (3) reduce visual time-sharing requirements so that attention can remain focused outside the cockpit or upon a primary flight display for longer periods of time. For example, the selection and tuning of radio frequencies, a procedure that often interferes with other essential manual tasks, could be controlled by voice. Computer generated displays, including head-up displays, could be controlled by voice rather than by conventional keyboard, thus avoiding eye-hand coordination problems and allowing visual attention to remain focused on the display. In addition, whenever command sequences or numerical data entries must be (1) performed manually, (2) integrated with other manual tasks, and (3) executed rapidly, e.g., a Navy P-3 sensor operator's use of thumbwheel switches for numerical data entry, voice commands have the potential for providing faster and more accurate performance, as well as being less disruptive of (and less disrupted by) other manual task requirements [7].

Once selection of a specific ASR flight application has been made and all hardware and software requirements have been met, full-mission flight simulation testing is desirable prior to actual use in flight.

Several candidate ASR flight applications are being considered for in-flight evaluation of the Ames ASR Flight System. Three of these potential applications have been selected for discussion.

Navy P-C3 Orion Pilot Application

During certain phases of anti-submarine patrol flights, Navy P-3C aircraft must be flown at low altitude. While low altitude flight is maintained, the pilot must select and execute command functions using a 35-key keyset that is located to his right and slightly behind him. This keyset location requires that the pilot turn his body towards the keyset and direct his visual attention away from the outside visual scene and his primary flight instruments while using the keyset. Use of the Ames ASR Flight System, with a 34-command syntax-structured vocabulary consisting of 11 subsets ranging in size from 2 to 14 commands would provide a command capability duplicating that of the keyset and would increase the pilot's capability for rapid detection and correction of significant deviations from the desired flight conditions during this time-critical phase of flight.

Lt. Anthony Quartano, a Navy P-3C pilot, developed the procedure used for selection of the voice command vocabulary and designed the command syntax in collaboration with the authors of this paper.

The speech recognition system would be implemented in parallel with the pilot keyset in a P-3C aircraft for in-flight evaluation of voice command system performance; thus comparable speed and accuracy data could be collected by voice and by keyset, and access to normal keyset functions would be available at all times.

Navy P-3C Sensor Station Operator Application

Sensor Station 1 and 2 operators on board Navy P-3C aircraft use several different keysets and thumbwheel switch input sets to enter data into a digital computer. During some missions, the desired rate of information entry greatly exceeds the rate obtainable with current input devices. Several Navy P-3C sensor operators based at Naval Air Station, Moffett Field, CA, have collaborated with the authors of this paper to evaluate potential ASR sensor station applications. Five specific applications were selected that would reduce an operator's manual workload and permit more rapid execution of essential tasks. For example, use of voice commands rather than thumbwheel switches for numerical data entry would increase the maximum entry rate while reducing interference with other concurrent manual tasks. Implementation of voice command functions would be made in parallel with existing input hardware so that comparable in-flight data could be collected by voice and by the conventional manual input method, and to provide access to normal hardware functions at all times.

Sufficient space is available in the Ames ASR Flight System chassis to accommodate up to four speech input boards, allowing the system to simultaneously process information from four different users. These users may or may not share vocabularies and/or input functions. This multi-user capability is made possible by the microprocessor-based speech input boards which relieve the 1603A computer of a large percentage of the total processing load. Thus the Ames ASR Flight System could easily process information in parallel from both Sensor Stations 1 and 2 during flight.

Helicopter Pilot Application

Helicopters, with their high noise and vibration characteristics, provide the most challenging and perhaps one of the most deserving flight environments for ASR applications. Because the manual control burden in helicopters is typically quite heavy, the implementation of an on-board ASR system could provide substantial benefits for increasing the efficiency and safety of helicopter flight. For example, during helicopter missions where success may require operating the aircraft near its operational boundaries, e.g., in search, rescue, or evacuation missions, the pilot could benefit from knowing how close he is to exceeding critical system performance limits such as altitude, airspeed, or gross weight. Given that an appropriate computational capability exists on the aircraft, information such as hover ceilings, gross weight margins, engine performance parameters, and range-of-flight estimates could be presented to the pilot in response to his voice commands. The fact that in these situations the pilot is in a high-workload manual control environment with inherent time-stress recommends the application of speech technology. An ASR system would not only allow the pilot to conveniently query the computational system for specific critical information, but speech synthesis could also be used to minimize the need for diversion of attention to a conventional visual display.

At the present time, Ames Research Center has assumed the lead role for NASA helicopter research, and applications such as this are being actively pursued. Fortunately, a variety of helicopters will be available for flight research in the near future, and in-house avionics groups are working on developmental projects such as the example discussed above. Accordingly, a very exciting and rewarding future is anticipated for this kind of man-machine systems research.

REFERENCES

1. Wempe, T.E. Flight Management--Pilot Procedures and System Interfaces for the 1980's - 1990's. AIAA Paper No. 74-1297, Amer. Inst. Aero. and Astro. Life Sciences and Systems Conference, Arlington, Texas, November 1974.

2. Nilsson, N.J. Learning Machines. New York: McGraw-Hill, 1965.

3. Hillborn, E.H. Keyboard and Message Evaluation for Cockpit Input Data Link. Report No. DOT-TSC-FAA-71-21, Dept. of Transportation, Washington, D.C., 1971.

4. Coler, C.R. and Plummer, R.P. Development of a Computer Speech Recognition System for Flight Systems Applications. Aersopace Medical Association, Preprints of the 45th Annual Scientific Meeting, Washington, D.C., May 1974, Pp. 116-117.

5. Belson, J. The 21st-Century Flight Deck. Flight International, April 23, 1977, 111, Pp. 1118-1120.

6. Turn, R. Speech as a Man-Machine Communication Channel (Report No. P-5120). Santa Monica: Rand Corp., January 1974.

7. Plummer, R.P. and Coler, C.R. Speech as a Pilot Input Medium. Proceedings of the Thirteenth Annual Conference on Manual Control, Cambridge, Mass., June 1977, Pp. 460-462.

DISCUSSION

Dr. Edward Huff

Q: Jerry Wolf, BBN: In one of your slides, Clay, you presented some recognition results for the entire 100-word vocabulary and then for comparison the 10 digits. The 10 digits came out with a smaller recognition score than the entire 100-word vocabulary. That seems a little inconsistent to me. I don't see how the smaller vocabulary could have come out with a significantly worse score than the 100. Do you have any comments on why that happened?

A: Clay Coler: As I indicated when we went through the results, when individual words were recognized over the entire vocabulary, some words had quite high recognition accuracy. For example, the digit six was recognized almost as well in the overall processing as it was by subsetting. Other digits were quite poor, and in fact the digit five was the poorest of all 100 commands. I think that the answer is simply that the digits are relatively tough, no matter whether they are protected in a subset or left unprotected in the overall processing.

Q: Jerry Wolf: They are significantly worse than the recognizability of that whole 100. They were relatively distinguishable.

A: Clay Coler: They are certainly significantly worse than other 10-command vocabularies and we can see that when we artificially put subsets together in the overall processing.

Bob Plummer: Maybe it wasn't clear that the ten words were still recognized against the whole hundred, but he was showing you the result of those ten particular words, the digits. So, that's how thay could come out with a lower score.

Q: Mike Grady, Logicon: I have to say that the work the NASA has done on trying to assess recognition accuracy is probably the most extensive I've seen, and I really want to express to all of you appreciation for that. I do have some questions on the experimental design or whatever you would like to call it. I guess I'll address this to you, Bob. In your comparison with 20 speakers in the six different training recognition strategies that you used, did you use the same 20 people when you went through each of the strategies, and also did the speakers, when they were performing the experiment, get feedback on the results that they were getting? I might point out that the reason I'm asking this question is because we found that there is really a phenomenon of learning how to talk to the boss, and it's a thing that we've never looked at and I'm curious about your results.

A: Bob Plummer: The idea there was to examine recognition algorithms separately from the frontend or the initial digitization. So, all the speaker did was say the right word at the right time. We would then digitize that and record the digital version of the word on tape. We did not do recognition at all on line. So the speaker received absolutely no feedback as to whether he was right or wrong. I mean for better or for worse, he didn't get it. What that gives us may be slightly unrealistic in the sense that in an application he probably would get feedback. On the other hand, it means that we have absolute repeatability because we can just cram those digitized words into any different recognition algorithm that we might want to use. So we really were applying all the different algorithms to exactly the same set of digitized data.

Q: John Markel Signal Technology: You talk about having 10-word reference or twenty-five words. Would you describe your training scenario? Whether that's taken at one setting or spaced over time. How is that done?

A: Bob Plummer: These particular tests were all done in the following way. The subject would see on a display the word or digit he would be asked to pronounce at a given time. All these data were taken sequentially, so in his first session he would provide us with twenty-five training samples of each word, and we would do that by cycling through the vocabulary, so he didn't say "waypoint" 25 times and then "altitude" 25 times. He would cycle through 25 times. Then after perhaps a short rest, he would go right on into the recognition. This all took place in one day, and more or less continuously through time. So we weren't really facing problems of day to day variation. Although I guess actually that is not completely true. Some of the data collection did take place over several days. Let me ask Clay--how many days were involved?

Clay Coler: For the 10-word vocabulary, all data were collected in a single day. For the 100-command vocabulary, it was five days. The first day was the training data day.

Bob Plummer: O.K., so what I said about the one day was right for the 10 words. For the 100 words, it was over five days. Training on the first day and recognition on the next four.

Q: Marv Herscher, Threshold Technology: I noticed on the slides and tables of the digits subsets, I think three times they appeared. One originally and two as subset-in your test, and there seemed to be some inconsistency in terms of the accuracies that were posted among the three. Would you comment on how this comes about?

A: Bob Plummer: I think that might be related to the original question It might be better to have a look later, privately or in the published paper. The question has come up in the following way: We were recognizing some times perhaps against only ten words, and that's what we call subsetting. So there the 10 digits were recognized only against other digits. At other times, we were recognizing overall. The digits were recognized against all 100 words but we artificially extracted out of that the digit subset to see how well those 10 words did as a group when compared to all 100 words. So those two results might look kind of different, but in fact I think we're consistent with what happened overall.

Q: Marv Herscher: I don't understand because I thought you said that the second column you had on a couple of the tables were digit subsets against each other. A true subset is a subset of words recognized only against each other.

A: Bob Plummer: Right. It was done three ways. The first was only the digit vocabulary. Then there was the big vocabulary with the digits as a true subset. And a big vocabulary with digits versus everything. Would you like me to put the slide back up?

Marv Herscher: Well, I can talk to you about it later.