N93-72621
[redacted]

VOICE DATA ENTRY IN AIR TRAFFIC CONTROL

DONALD W. CONNOLLY

NATIONAL AVIATION FACILITIES EXPERIMENTAL CENTER
ATLANTIC CITY, NEW JERSEY

59-32
176345

BACKGROUND

The introduction of large-scale automation into civil air traffic control is relatively recent. Until 1970 there were only a very few isolated and largely developmental installations. Since 1970 all 20 of the enroute control centers and 64 of the major terminal control facilities have had large, computer-based systems installed and commissioned for operation. These systems function in many ways analogously to military command, control, information and communication systems and they are all directly or indirectly interconnected.

The whole air traffic control complex is basically a cooperative surveillance and control operation. Its functioning depends on many elements, not the least of which is complete, correct and up-to-date flight plan and flight progress information. The execution of flight plans precisely as filed in advance, however, is more exceptional than routine. Even "standard" airline plans for scheduled flights are subject to change before departure as well as while enroute due to many factors, most notably weather and wind conditions. The air traffic control specialist, of course, is the principal point of direct contact between the traffic management system and the traffic itself and is a major conduit of information between them. While the function of the system is to maintain and provide vital information to the controller, he or she in turn has the task of supplying a substantial quantity of information to the system.

Automation has altered a number of task elements of the job of the air traffic controller. In most instances these changes have been in the direction of improved quality, efficiency and simplicity though they have by no means diminished the complexity or responsibility of the job of traffic controller. While some types of workload have been reduced or nearly eliminated (visual/manual tracking, maintenance of identification, acquisition/maintenance of altitude information) new tasks have been added. Perhaps the most significant and onerous of the latter is that of manual entry of flight data and system commands and queries. As in many computer-based operations the "language" used for entry or query is an abbreviated, partially mnemonic, coded language. Even so, the key-entry workload and its potential for distraction and interference with the main stream of the controllers' task, remains high. In peak traffic hour, for example, an enroute traffic controller will

commonly find it necessary to enter messages into the system computer which aggregate to 700 or more single keystrikes.

Human factors and air traffic control specialists at the National Aviation Facilities Experimental Center have recognized and been concerned with information transfer problems at the controller-computer interface throughout the system development process. A major interest, of course, has been the area of data entry and system control. Emerging technologies of information presentation and data input are continually reviewed and promising techniques experimentally investigated. A number of variations of the "touch" or "menu-select" principle of "chunk" data entry (as versus character-by-character message composition) have been tried in laboratory and simulation experiments, for example. We have been aware of, and following, the development of spoken word recognition technology since at least 1971. It was not until the middle of 1975, however, that we were able to secure the approval and resources necessary to undertake in-service exploration of the applications of word recognition (and, by some logical and temporal extension, speech understanding) in the field of air traffic management. Thus far the magnitude of effort underway in the Federal Aviation Administration has been rather small (one scientist with part-time aid of one technician and one programmer) and has been directed toward application, adaptation and modification of speech recognition technology rather than development of the technology itself.

PROGRESS

Introduction

In May 1975, a basic Threshold Technology, Inc., model VIP-100 was acquired for use in a series of word-recognition applicability studies. This equipment included an ASR-33 Teletype, a NOVA 2 minicomputer with 16K of core memory, the Threshold digitizer and a three-transport cassette tape unit. A Tektronix model 4012 CRT/keyboard computer terminal was added as the basic output device. At various times since 1975 additional hardware has been secured, including a 10 megabyte disk store, a Digital Equipment Corp. DECwriter, 16K words of core memory and an in-house designed and fabricated voice digitizer whose uses and potential uses will be described below under post FY-77 efforts. The equipment of the Voice Entry Laboratory is shown schematically in Figure 1.

Several of the keyboard data entry "languages" of the National Airspace System were tabulated and analyzed. There are two such languages in regular and extensive use in the semi-automated enroute traffic control centers of the agency which produce daily hundreds of thousands of messages requiring millions of keystrikes. There are a number of

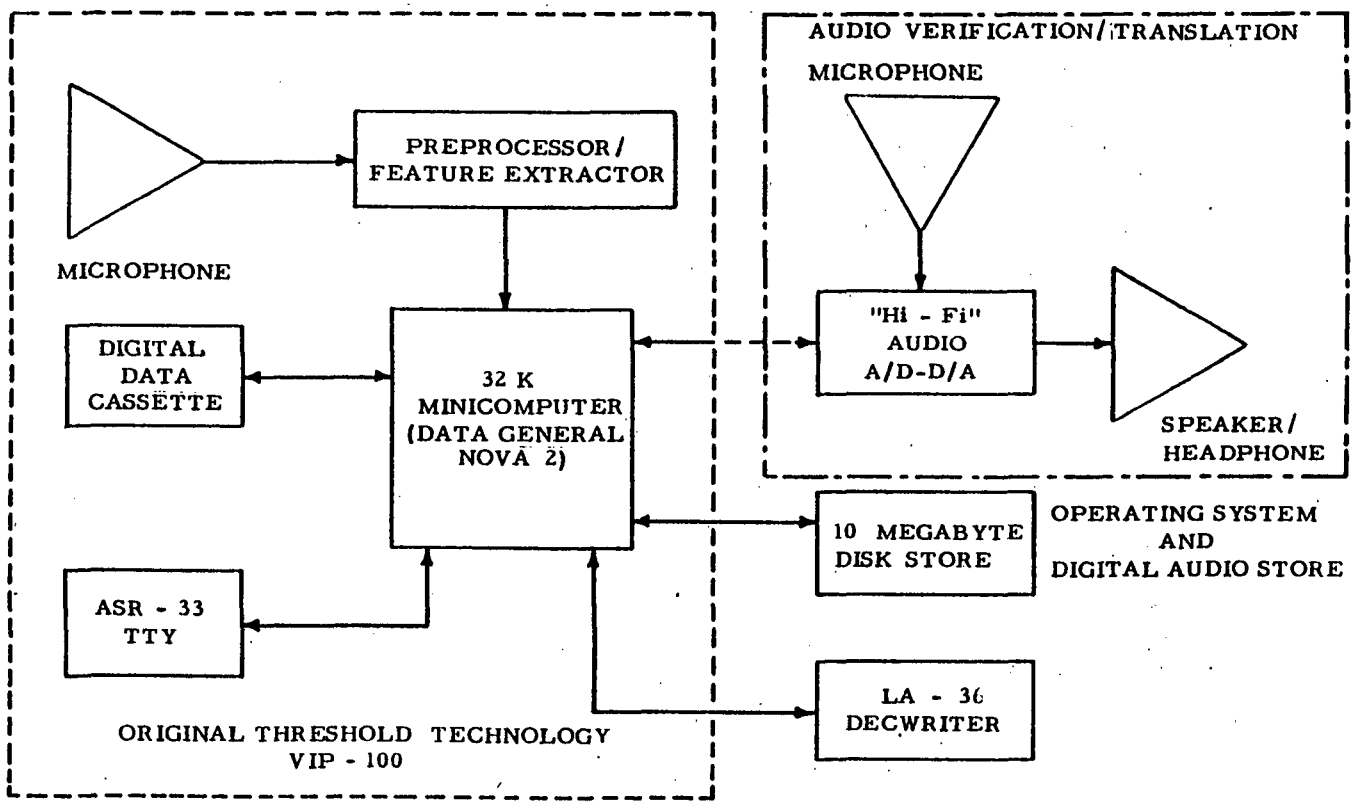

FIGURE 1
FAA/NAFEC VOICE ENTRY LABORATORY

other entry languages in the system (e.g. control tower cab, terminal radar control facility, flight service station, etc.) which are either not as burdensome or distracting, or not as complex and voluminous in use, or both, but which are also likely candidates for application of word recognition technology. The key language which was chosen as the test vehicle was that used by the non-radar or flight data controllers in enroute control centers. The structure and vocabulary of this lang-usage may be found in Tables 1 and 2. This particular language was selected for a number of reasons. In the first place, it is one of the more complex languages in use. The total repertoire of possible messages is larger than that of any of the other key languages used by personnel engaged in the active control of traffic. Finally, the key-entry workload at this operational position is the largest in total volume in the system. Thus, a very difficult application was undertaken for investigation right at the outset. The theory behind this choice was that (a) it appeared highly likely, given the state of the word recognition art, that this application would be practical and cost/beneficial and that, a fortiori, less complex, less difficult applications would yield to the same approach with zero or minimum additional research and development effort or that (b) many or most of the relevant questions for the lesser applications would be answered in the course of attacking the greater, even if the present state of technology did not prove practical for this particular application.

Initial Experiments

The language chosen for test was found to include a total of 24* basic types of messages. Of these, 15 types of messages encompass 96 percent of all messages actually entered in operation. In addition, these 15 message types include all of those occurring with a frequency of one in a hundred or greater. The first element of every message is the message type. It was also found that, in most cases, the type of message must be followed by the identity of the flight data file (flight plan) to which the entry applies. Furthermore, of the four means of identifying a flight, the one most commonly employed was the three-decimal-digit computer identity number assigned to every flight. Thus, the second element of most spoken messages could be assembled from a word list consisting only of digits plus two or three control words (such as "erase" for restarting the whole entry and "backspace" for changing the last digit.)

*An additional seven types of message (covering "conflict alert" entries) have since been added. This, based on experience to date, should not cause any special difficulty.

TABLE 1

VOICE DATA ENTRY: D-CONTROLLER LANGUAGE STRUCTURE

KIND OF MESSAGE				
AMEND	3 DIG IDENT	DATA FIELD NAME	DATA ENTRY FOR FIELD	GO
CORRECTION		"	"	"

DATA FIELD NAMES	VOICE ENTRIES REQUIRED
Type	See Below
Beacon Code	4 Octal Digits
Speed	3 Decimal Digits
Fix	Place Name
Time	4 Decimal Digits
Altitude	3 Decimal Digits
Qualifier	See List of Qualifiers and Note 2
Ident	7 Alphanumerics (decimal)

Note 1: After a "field name" and appropriate entries for that field have been entered the system will accept another field name (plus proper entries) and yet another etc. without limit OR it will accept an ERASE command, a BACKSPACE command or a GO (ENTER) command.

FOR "TYPE" ENTRIES, ALWAYS SAY:

	MFG NAME,	2 or 3 A/N,	Name a Qualifier
or	MILITARY,	4 A/N,	" " "
or	GENERAL,	4 A/N,	" " "

IF YOU SAY:	YOU'LL SEE:	THEN SAY:
Boeing	B	3 A/N e.g. 707
British	BA	2 A/N e.g. 11
Vickers	VC	2 A/N e.g. 10
Lockheed	L	3 A/N e.g. 011
Nord	N	3 A/N e.g. 026
Dehavilland	DH	2 A/N e.g. C6
Douglas	DC	2 A/N e.g. 10
Military	---	4 A/N e.g. C131
General	---	4 A/N e.g. PA13

TABLE 1 (Continued)

TO ENTER A "TYPE, YOU MUST ALWAYS ADD ONE OF THE EQUIPMENT QUALIFIERS:

IF YOU SAY:	YOU'LL SEE
Discrete	/U
DiscreteDME	/A
DME	/D
Nondiscrete	/T
NondiscreteDME	/B
Transponder	/X
TransponderDME	/L
TACAN	/M
TACAN64	/N
TACANDiscrete	/P

Note 2: If you wish to enter an amendment to the QUALIFIER part of the "type" field alone, you need only name the data field "QUALIFIER" then name one of the qualifiers above.

FINALLY, YOU MAY SAY "GO" (to ENTER), BACKSPACE (if you wish to change or correct an error of entry or of recognition) or "ERASE", OR, YOU MAY NAME ANOTHER DATA FIELD AND CONTINUE AS BEFORE.

KIND OF MESSAGE	SEQUENCE		
REPORTALTITUDE	3 DIG IDENT,	3 DECIMAL DIG (ALT),	GO
DISCRETECODE	"	4 OCTAL DIG (CODE),	GO

These are shorthand messages requiring only the KIND name, the track I.D. and the data to be entered. It is realized that discrete codes are often assigned automatically upon entry request in NAS. The voice entry here is for test purposes.

TABLE 1 (Continued)

KIND OF MESSAGE	SEQUENCE	
DROPTRACK	3 DIG IDENT,	GO
PRINTSTRIP	"	"
ACCEPTHANDOFF	"	"
READOUT	"	"
CANCEL	"	"

These messages are all identical excpet for the first word, the kind of message.

KIND OF MESSAGE	SEQUENCE			
DEPARTURE	3 DIG IDENT,	4 DEC DIG (TIME)	NAME (FIX)	GO
HOLD	"	4 DIG (TIME)	NAME (FIX)	GO
RELEASE	"	4 DIG (TIME)		GC
TRANSMIT	"	NAME (FIX)		GO

These messages require entry of a four digit time, or a one word place name (FIX) or both in addition to the message kind and the identity of the flight.

KIND OF MESSAGE	SEQUENCE		
WEATHER	NAME(FIX)		GC
HANDOFF	2 DIG (SECTOR)	3 DIG IDENT	GC

These and CORRECTION (above) are the only kinds of messages that are not immediately followed by identity.

TABLE 2

VOICE DATA ENTRY: D-CONTROLLER VOCABULARY

WORD NO.	UTTERANCE	PRINT WORD DISPLAYED
DIGITS		
0	ZERO	0
1	ONE	1
2	TWO	2
3	THREE	3
4	FOUR	4
5	FIVE	5
6	SIX	6
7	SEVEN	7
8	EIGHT	8
9	NINER	9
CONTROL WORDS (SEE ALSO #102 ERASE)		
10	BACKSPACE	
11	GO	(ENTER)
MESSAGE TYPES		
12	AMEND	AM
13	CANCEL	CN
14	CORRECTION	CR
15	DEPARTURE	DM
16	DISCRETECODE	DQ
17	READOUT	FR
18	ACCEPTHANDOFF	HO
19	HANDOFF	HO
20	DROPTRACK	RS
21	PRINTSTRIP	SR
22	HOLD	HM
23	RELEASE	HM
24	REPORTALTITUDE	RA
25	WEATHER	WR
26	TRANSMIT	XM
FLIGHT DATA FIELD NAMES		
27	TYPE	03
28	QUALIFIER	03
29	BEACONCODE	04
30	SPEED	05
31	FIX	06
32	TIME	07
33	ALTITUDE	08
34	IDENT	02

TABLE 2 (Continued)

FIXES

35	WILLIAMSPORT	IPT
36	SELINGSGROVE	SEG
37	MILTON	MIP
38	HAZELTON	HZL
39	WILKESBARRE	AVP
40	EASTTEXAS	ETX
41	LAKEHENRY	LHY
42	TOBYHANNA	TSD
43	ALLENTOWN	ABE
44	STILLWATER	STW
45	BENTON	7QB
46	SWEETVALLEY	7EV
47	LOPEZ	7LE
48	SNYDERS	7YX
49	SLATINGTON	7ZO
50	WHITEHAVEN	9WT
51	RESORT	9ZT
52	PENNWELL	7PW
53	HUGUENOT	HUO
54	SOLBERG	SBJ
55	FREELAND	7FE

AIRCRAFT TYPE NAMES

56	BOEING	B
57	DOUGLAS	DC
58	LOCKHEED	L
59	CONVAIR	C
60	VICKERS	VC
61	NORD	N
62	BRITISH	BA
63	GENERAL	-
64	MILITARY	-
65	DEHAVILLAND	DH

PHONETIC ALPHA

66	ALPHA	A
67	BRAVO	B
68	CHARLIE	C
69	DELTA	D
70	ECHO	E
71	FOXTROT	F

TABLE 2 (Continued)

PHONETIC ALPHA (Continued)

72	GOLF	G
73	HOTEL	H
74	INDIA	I
75	JULIET	J
76	KILO	K
77	LIMA	L
78	MIKE	M
79	NOVEMBER	N
80	OSCAR	O
81	PAPA	P
82	QUEBEC	Q
83	ROMEO	R
84	SIERRA	S
85	TANGO	T
86	UNIFORM	U
87	VICTOR	V
88	WHISKEY	W
89	XRAY	X
90	YANKEE	Y
91	ZULU	Z

"QUALIFIERS"*

92	DISCRETE	/U
93	DISCRETE DME	/A
94	DME	/D
95	NONDISCRETE	/T
96	NONDISCRETE DME	/B
97	TRANSPONDER	/X
98	TRANSPONDER DME	/L
99	TACAN	/M
100	TACAN 64	/N
101	TACAN DISCRETE	/P

*These expressions are to be said as all one word such as "discrete dee em ee", even though printed here and on the training display as separate words.

CONTROL WORD
(SEE ALSO #10 BACKSPACE AND #11 GO)

102	ERASE	Erases Entry

The second element of some types of messages (e.g., weather information retrieval) and third or fourth element of other messages (e.g., early handoff to a terminal; hold message) is a location identifier or geographic "fix." The keyboard codes for these place names are not always mnemonic (e.g., Benton is coded 7QB) but the place names themselves are easily spoken. No attempt was made to survey all possible fix-names; however, the list included for one sector in the New York ARTCC, all VOR's, all intersections, and all terminals; in short, all the fixes normally required at the position as elements of key-entry messages.

Two types of messages (flight plan amendment and correction thereof) require identification or naming of a flight plan data field (e.g., assigned altitude; speed). Eight of these data fields account for the vast majority of modifications entered and the field content or substantive data most commonly consist of digits.

Certain types of entries or, more precisely, parts of messages currently made with keyboards basically exist only in coded, nonverbal or partially nonverbal form. Consider the aircraft identity N1009Y (tail number). The most convenient way to make such an entry might still be via keyboard. However, an "all purpose" subvocabulary consisting of all of the digits plus the phonetic alphabet (which is part of the linguistic stock-in-trade of the traffic controller) were made a part of the total vocabulary of the voice data entry language for the purpose of making the comparatively fewer and rarer entries not already encompassed by the word lists described above.

These subvocabularies, plus a short list of commercial aircraft types and the list of relevant avionics equipments (or type "Qualifiers"), make up the whole vocabulary as currently constituted. The vocabulary and syntax of the language, as previously noted, are included here as an appendix.

The first experiments which were conducted were intended to establish the basic recognition performance of the VIP-100 word recognition package with three of the subvocabularies discussed above, namely the 15 message types word list, the 21 fix names list, and the 10 digits (plus "erase" and "backspace") list. Each of the lists, separately was expanded into a pseudo random assembly in which each member of the list appeared 10 times. Thus the "reading list" for message types was 150 "words" long, that for "digits" 120 words, and for fixes, 210 words. Each speaker then "trained" the word recognizer by speaking each expression (some, as may be seen in the appendix, were composites or phrases spoken without internal pauses) 10 times. This resulted in composite digital images of the way the speaker speaks each of that particular list of words. These reference images were then written on cassette

tape for later reuse. Following the initial "training" session, each speaker reads the random list described above on 10 separate occasions in the case of message types and fixes, 5 sessions for the digits list. Data were automatically collected during each test session on the number of times each word was correctly recognized, the number of times incorrectly recognized, the average closeness of match between the spoken entry and the best and second-best choice among the reference images (i.e., the training images), and the duration of the spoken expression. Each subject, over a period of several days to several weeks, spoke (for recognition testing) each word in each of the subvocabularies 100 times for the types and fixes and 50 times for the digits. The principal purpose of testing digits at all was to ascertain whether our sample of speakers produced the order of recognition accuaracy for digits which is commonly found using this word recognition equipment.

Initial Results

A total of 12 speakers served as test subjects for Phase I. Nine were journeyman air traffic control specialists with extensive experience in the National Airspace System Enroute Test Facility. Three were non-controllers, two female and one male. No differences were found that could be attributed to either profession or gender. One group of 11 of these speakers served as subjects for the message types (nine male, two female) and another group of 11 from the same pool of speakers served for the other two word lists. Each entry in the Recognition Accuracy column in Table 3 is based on a total of 1,100 entries of the word for types and fixes and 550 for digits; thus, each is considered quite reliable.

Most of the words in the three subvocabularies of this language were recognized with an accuracy of 98 percent or better. This figure does not include a rejection rate that also averaged about 1 percent (i.e., the utterance was not recognized as acceptably close to any of the reference images of the list at all). The error data are considered more critical, since the speaker's attention can be called easily to a "rejection" while misrecognition must be detected by the speaker himself.

To take one example of a rough comparison between spoken and key entry, consider the list of geographic fixes. Key entry of each requires striking three keys in an "artificial language." Thus, our 11 speakers made entry of each of 21 fixes 100 times or a total of 23,100 spoken entries. Overall accuracy of recognition was 99 percent

TABLE 3

VOICE ENTRY SUBVOCABULARIES

D-CONTROLLER MESSAGE TYPES

WORD	KEY CODE	RECOGNITION ACCURACY*
AMEND	AM	97.73
CANCEL	CN	94.36
CORRECTION	CR	99.73
DEPARTURE	DM	98.18
DISCRETECODE	DQ	99.91
READOUT	FR	99.91
ACCEPTHANDOFF	AHO	97.55
HANDOFF	HO	99.18
DROPTRACK	DROP	99.64
PRINTSTRIP	SR	98.82
HOLD	HM	99.82
RELEASE	REL	100.00
REPORTALTITUDE	RA	98.09
WEATHER	WR	99.36
TRANSMIT	XM	97.09
		(98.62 Overall)

DIGITS (IDENTITIES, SECTORS, DATA)

WORD	KEY CODE	RECOGNITION ACCURACY**
ZERO	0	99.82
ONE	1	97.82
TWO	2	100.00
THREE	3	99.82
FOUR	4	99.09
FIVE	5	99.09
SIX	6	100.00
SEVEN	7	99.64
EIGHT	8	96.36
NINE	9	98.91
ERASE	--	99.64
BACKSPACE	--	100.00
		(99.18 Overall)

*Each entry based on 1,100 spoken inputs
**Each entry based on 550 spoken inputs

TABLE 3 (Continued)

FIX NAMES

WORD	KEY CODE	RECOGNITION ACCURACY*
WILLIAMSPORT	IPT	98.82
SELLINGSGROVE	SEG	98.91
MILTON	MIP	96.45
HAZELTON	HZL	97.45
WILKESBARRE	AVP	99.73
EASTTEXAS	EXT	99.91
LAKEHENRY	LHY	99.91
TOBYHANNA	TSD	99.45
ALLENTOWN	ABE	99.82
STILLWATER	STW	99.82
BENTON	7QB	98.64
SWEETVALLEY	7EV	99.91
LOPEZ	7LE	99.82
SNYDERS	7YX	99.73
SLATINGTON	7ZO	99.36
WHITEHAVEN	9WT	97.27
RESORT	9ZT	99.27
PENNWELL	7PW	99.73
HUGUENOT	HUO	98.45
SOLBERG	SBJ	99.09
FREELAND	7FE	99.18
		(98.99 Overall)

*Each entry based on 1,100 spoken inputs

and approximately 1 percent of entries were rejected entirely. Key entry of these same characters would require 69,300 keystrikes with essentially no protection whatever from single-key errors, while each voice entry results in display of the whole three-character code for the fix which seems more susceptible of error detection than one or more misstruck keys. The time involved in the two entry methods seems indistinguishable. We did not collect accuracy data on key entry of fixes but this will be an integral part of later experimentation wherein voice versus keyboard entry of complete messages will be tested.

Follow-on Reliability Studies

While the recognition accuracy data for the subvocabularies of this language were impressive overall, two major considerations impelled us to seek methods of improvement. In the first place, it must be remembered that the "user" here is the air traffic controller and the principal aim of voice data entry is reduction of distraction from his or her main concern, namely continuous observation and management of the dynamic four-dimensional traffic situation. It is thus essential that detection and correction of data entry errors be brought to some irreducible minimum. The second problem is that of individual differences in recognition accuracy from speaker to speaker. While precision and clarity of speech are of the essence in air traffic control, some controllers necessarily will speak with greater uniformity than others. Thus, while the overall recognition error rate for the message types subvocabulary was less than 1.5 percent, individual speaker error rates ranged from less than 0.1 percent to nearly 7 percent. With the "digits" subvocabulary, the overall average error was less than 1 percent while the range was from zero to 2.3 percent. Similar results were obtained for the subvocabulary of fix names.

It was decided, therefore, to investigate means of error reduction and/or error correction which might be applied to the basic VIP-100 recognition algorithm. We consulted with Dr. Breaux of the Naval Training Equipment Center regarding some of the recognition subroutines that he had developed for increasing recognition accuracy in his application in the ground controlled approach trainer. These, as well as a variation of the same general concept which was developed for us by Mr. Cox of Threshold Technology were experimentally tried with the non-radar controller data entry language with which we have been working. The net result, despite manipulation of the parameters of these routines, was either an increase in rejected inputs or an increase in the error rate or both. In retrospect this should not have been surprising, since the logic of these techniques is directed principally to the solution of the recognition problem where the input utterances are relatively long and largely identical with the exception of a single element. For example, the expressions "slightly (above/below) glidepath" can be differentiated

with greater accuracy if both the reference and the input images are pared down to only those parts which are non-identical and a "second look" taken at the correspondences. This precise situation did not obtain in the word lists used here. The more common type of problem encountered was confusion of some of the pairs of words within a subvocabulary. The words "transmit" and "printstrip" in the message types list and the words "Williamsport" and "Resort" in the fix names list were among the frequent confusions. Oddly enough, even though the expression "nine" (instead of "niner") was used in the digits word list, and nearly all errors involved the five/nine and nine/five confusions, a very high order of accuracy was obtained for both words.

In the course of trying out various alternative decision subroutines for error reduction and in re-examining our original detailed data we were struck by some interesting features of the word durations. For every utterance in the original tests we recorded the word numbers and correlations for the best and second-best matches and the duration (i.e., number of audio samples) of the input utterance. In the course of time normalization of utterances, we had been discarding this information after use. It was an interesting curiosity of our subvocabularies that some of the confusions that were common (such as Williamsport/Resort and fix/backspace/erase) were quite reliably distinguishable on the basis of utterance duration. In the course of investigating the utility of this phenomenon in turn (we started collecting utterance duration data during the "training" or reference array construction mode of operation) we further discovered that there were systematic differences in utterance duration during "training" as versus "recognition." The average duration of the utterance spoken repetitively during training frequently differed from the average duration of the same utterance spoken in a pseudorandom sequence. Since the durations differed under the two conditions, it was hypothesized that the correlations obtained in recognition would necessarily suffer.

The software was then modified in two ways. First, training was changed so that the speaker was presented with a pseudorandom prompting list. He or she did not simply repeat each word in the list in times in succession, but rather in times within the same list but seldom or never the same word twice in succession and in an unpredictable order. At the same time, the average duration of each word as well as the shortest and longest obtained during training were recorded and made a part of the reference information. The recognition decision algorithm was changed to make use of the duration data. The basic logic is as follows:

1. The input utterance is digitized, time normalized and its duration is noted.

2. The normalized feature array is compared with reference arrays for all words in the subvocabulary and the routine returns with the correlations for the best and second-best matches.

3. If the correlations differ by more than 40, the best match is selected as correct.

4. If the correlations differ by 40 or less, the input utterance duration is compared to the average (during training) duration for the first and second choice words unless the latter two durations themselves differ by less than 30 samples.

5. If the duration of the instant input is closer to the reference duration of the first-choice word, it is accepted as correct.

6. If the duration of the utterance is closer to that of the second-choice word, the utterance is rejected.

7. If the two reference durations differ by 30 samples or less, the test is not made and the first choice word is accepted as correct.

In addition to these changes in the training and recognition algorithms, we added a "tuneup" mode of operation to the basic program. In this mode of operation, the speaker puts on and adjusts the headset, adjusts the input volume setting and then starts reading the words in the particular subvocabulary. The recognition decision word is displayed on the Tektronix terminal CRT and just below it, the duration in samples of the utterance just made and the average duration of the first-choice (or recognition decision) word. If the two durations are not reasonably close (i.e., differ by more than 10 or 15 samples) for several of the words, even when repeated several times, then the headset placement and volume setting are rechecked. This "tuneup" mode is also useful for checking the effects of a cold or other speech altering event and the need for "retraining" specific words.

Follow-on Results

Having made new training data by the pseudorandom repetition method, two of the "better" (i.e., higher overall recognition accuracy) and two of the "poorer" speakers were retested on the three subvocabularies previously used. With only one exception (fix names for one of the "better" subjects) the difference between the average duration of utterance in the training or reference data and the average duration of the same utterances under recognition conditions decreased substantially. With another similar exception, the average correlations of input utterances increased. That is to say, the quality of the matches between the inputs and their reference images, on the whole, improved. As might be expected, overall errors of recognition were reduced. The percentage error across all speakers and all three word lists went from 1.0 down

to .35 percent. The percentage of rejects, somewhat surprisingly, went from 1.3 down to .8 percent. This last is surprising because it was expected that the use of duration information in the recognition decision logic would tend to increase the reject rate by rejecting some doubtful, atypical but correctly recognized (on the basis fo correlation alone) spoken inputs. This was a trade we were willing to make, namely, the exchange of rejects for errors. The "cure" for a rejected entry is simple: Say it again. The cure for an error is another story entirely.. Thus it would seem that the modified training routine alone solved most of the problem we sought to solve. In addition to this effect, the duration test in the decision logic only slightly increased the reject rate for two of the speakers on the list of fix names while the error rate for both was reduced to zero. Indications are, overall, that use of this additional information will convert a portion of the potential errors to rejects for some talkers.

Recognition reliability or error rate improved for both the "poorer" and the "better" talkers on all three subvocabularies with only two exceptions wherein it simply remained the same. In one of these two cases the error rate was zero under the original test conditions and, obviously, could not have been improved in any event. The improvements for the "poorer" talkers were not uniformly dramatic but they were very impressive in most cases.

It must be admitted that in the follow-on studies reported here we were proceeding on a "pilot-study" or "cut-and-try" basis until the very end. Thus, the final results noted just above are accounted for by a combination of variables. The training procedure was changed, the "tune-up" feature was added and the decision logic was modified. In addition, there may have been some unknown quantity of "Hawthorne Effect" upon the "poorer" talkers who worked closely with the experimentors through the cut-and-try phase of experimentation. The "acid test" of the objective changes should properly be made with a new sample of subjects but it does not seem likely at the present time that we will be given the time and resources necessary to accomplish this. On the whole, however, we feel that we have substantially realized our goal which was reduction of recognition error as close to the vanishing point as possible given the technology at hand. We believe that perhaps three to five errors of recognition in a thousand entries is a tolerable level within which to pursue further the applications which we have in mind. We fully expect that this error level will increase to some degree under conditions of lengthy message assembly as distinguished from subvocabulary testing. It remains to be seen how much it increases and what the subsequent ramifications (in user acceptability, for example) of such an increase may be.

Other Findings

A great deal of ancillary but, from the standpoint of our potential applications, relevant and important information was also obtained. Our data from the initial reliability testing were studied for information on questions about speaker learning or familiarization effects, effects of such factors as colds and allergies on the recognizability of speech, effects of different types of microphones, and of precision placement of microphones.

In the matter of user familiarization and training, several important observations were made. During the test series for each speaker with each word list, recognition accuracy and "rejection" data were processed not less often than after every second session. As a rule, in the event that any individual word was either erroneously recognized two or more times or rejected as unrecognizable two or more times, a new set of "training" data was made for that word (and, in the case of errors, for the word with which it was confused if the confusion was consistently between the same two words). Thus, as recognition testing proceeded, the quality of the reference images or "training data" for some of the words in each list for some of the speakers was progressively refined. This does not mean that a great deal of retraining was done. A number of the speakers never needed to "retrain" any of the words in any of the lists at all. On the average, each speaker needed to retrain one word one time for the list of fixes, for example. Some speakers needed to retrain more words than others and some of the words and word pairs were more troublesome than others. See, for example, the fixes Milton and Benton in the list of fix-names. Attempts by some speakers to adopt an extraordinary (for them) pronunciation or emphasis in an attempt to improve recognition of a word were disastrous. Habitual or "natural" expression of the utterances is vital to accuracy of recognition. The modified training routine and our version of a "second look" in the decision logic (plus the long familiarity of the subject speakers by the time these were tested) reduced the retaining requirements to nearly zero.

Colds and allergies which affect the characteristics of speech were found to deteriorate recognition quality. However, for two of three speakers who among them contracted three head colds and one allergy during the test series, no serious problems were encountered. For these two speakers, it was necessary to retrain only a few of the words in the list to recover the near-perfect recognition previously found. One speaker, indeed, contracted a second cold after several weeks. It was only necessary to read in to the system the training data modified for the first cold in order to achieve the same recognition quality as produced by the "normal speech" training data. The third speaker, however, despite major efforts at retraining specific words was unable to

regain a high recognition accuracy while the cold persisted. It should be noted that the overall data for recognition of message-type entries which has already been discussed includes the error data from this speaker which accounts for approximately half the total errors encountered with this particular subvocabulary. When this speaker was not suffering from a serious cold, his results were quite comparable to those of other speakers.

Retests were also run with most of the original twelve speakers using the last (and best) set of training, or reference, data recorded during the initial reliability testing phase. Retests were made after approximately 3 months and again after approximately 6 months following the last of the original test series. Both accuracy and reject results were almost identical to those found in the initial test series.

Finally, microphone quality and placement were found to be factors of influence. While fully systematic testing of these variables was not conducted, three different (but all "noise canceling") microphone types with four different mountings (one hand-held, three headset or headband) were employed at various times. The hand-held microphone was used by three of the speakers during the testing of the 15-word message-type list and accounts, in part, for the slightly lower overall accuracy rate found for that list than for the others. Careless, inconsistent, or unusual placement of microphones (e.g., at or below chin height, more than an inch from the corner of the mouth in the horizontal plane) immediately appears in a high reject rate because of loss of signal strength and can quickly be corrected by the user. Throat-type microphones were not tested but might be worthy of trial. The microphone used by all but one subject for the "digits" subvocabulary is directly substitutable in existing air traffic control operations for the carbon-type microphones required by the communications systems employed today. This microphone produced excellent results. Some further testing using actual carbon microphones belonging to field operations is planned.

FY-78 PLANS

Over the next year we plan to complete at least some of the original overall experimental application plan which includes:

1. Experimental data collection in a series of "keyboard vs. voice entry" experiments. One of these is expected to be a laboratory, baseline establishment type of effort. A number of operators will simply make entry of a large number of traffic control computer input messages by both methods. This will provide a solid basis for: (a) Assessing the absolute and comparative reliability and efficiency (as

well as "user acceptability") of word recognition technology in this type of application, and (b) Assessing the subsequent effects of task-induced stress, the mixture of air-ground-air voice communication with ground-ground (controller/computer) communication by voice and other factors as yet unpredictable.

2. Addition of voice-feedback or audio verification to the system. Preliminary results indicate the possibility of no real gain in speed of entry but hard figures on accuracy of key vs. voice are not yet in hand. The principal gain we envision is reduction of distraction, a significant safety factor. Audio message verification may be an essential element of this last.

3. Field testing. Everything else being equal, a miniaturized (micro-computer) version of the "final" design will be (we hope) brought to a number of operational control facilities for field operator evaluation.

In subsequent years we will possibly experiment with language translation. The audio feedback technique we plan to use (which has, by the way, already been built in-house and is being tested) is based on digitization of real speech on a high sampling rate, limited feature count basis, similar to that used by long line telephone systems. We store these digitized images (practical at present for only a very limited language), concatenate and reconstitute them. The voice output quality is excellent and there is no problem of synthesis, especially of multiple natural languages. There is no raw synthesis, in fact, merely re-conversion from digital to analogue.

We have most of the software and nearly all of the hardware necessary to undertake these activities. The principal deficiency is people just now--the time of the principal investigator and the availability of subject/talkers.

POST FY-77 TECHNOLOGY REQUIREMENTS

For the applications which we envision (and probably for many others) the following seem to be the "breakthroughs" needed to absolutely assure the future of voice technology in real-time, interactive command and control:

1. In the "word recognition" area (as distinguished from speech understanding), better word boundary detection. Maybe this really means limited SUS--for three and four digit numbers, for example. Many of our operational key-languages consist largely of numerical entries. Speed is important, here of course. None of our applications can wait seconds, much less minutes, for rendition of composite or even continuously uttered two to four digit expressions.

2. "Better" microphones or "less sensitive" (but equally accurate) digitization or both. What is meant here is a solution to the problems of speech-type noise, the necessity of precision microphone placement, the necessity of "calibration" of digitizers to microphones, some of the as yet unknown problems of speaker stress and similar accuracy-lowering factors. The solutions here might lie solely or principally in software, though possibly "adaptive" or "intelligent" hardware or some combination.

3. Continued improvement of the "many speaker" capabilities of SUS's. While the applications we envision can get by with pretrained systems (and, indeed, necessitate the precision and speed of the single-speaker, word-recognition technology as versus those of the present state of SUS technology) since we always know who the operator is going to be, these improvements would certainly be in the "nice to have" class. For some civil aviation applications, this characteristic is virtually an essential requirement. Applications here include general aviation pilot briefing and audio flight-plan entry.

BIOGRAPHICAL SKETCH

Donald W. Connolly

Engineering Research Psychologist
National Aviation Facilities Experimental Center
Federal Aviation Administration

1967 - Present:	Primary responsibilities in human factor engineering of air traffic control ground environment - ATC facilities. Research, development, test and evaluation in man/machine interface technology - displays, controls, workplaces and operating procedures.
1952 - 1967:	Previous experience in human engineering of Air Force, Army and Navy command and control systems and in basic and applied research in human factors with New York University College of Engineering and as a civilian with the U. S. Air Force.
Education:	B. S. '50, M.A. '52, Ph.D. '57, Fordham College/ University, New York, N.Y., Experimental Psychology.

C-3

DISCUSSION

Donald W. Connolly

Q: Mike Curran: You envisioned a success rate, 995 or 997 out of a thousand tries. That's pretty good.

A. I know that as we add factors of complication (stress factors), in other words, full message composition vs. just plain vocabulary testing, this rate will deteriorate. There is the question, for instance, of the task induced stress of actual control of traffic in a tense situation. It may or may not be a problem. But it probably will be a problem. One of the things I have discovered myself is that if you sit there and talk for four hours something happens to your voice. You may have to compensate for that too in some way.

Q: Mike Curran: Don, I didn't ask you the question yet. I'm giving you that success rate of 999 out of a thousand presuming we can ever get there. In your application do you still see a need for verification before the entry? Do you see the approach as a non-verified entry?

A: I think probably not, in the practical day-to-day operational sense. For instance, data entry, an error in the entry of data into the flight plan file is not fatal. It's lots of things, but it is not ordinarily fatal. Errors in communication between the controller and the pilot can be very serious. On the other hand, I see some kind of verification at least nice to have. Something which perhaps could be turned on and off, I don't know. At least for the beginning, absolutely essential. We have a redundancy in everything we do now. We still have the paper flight progress strips that they used in 1939, all racked up, just in case.

Q: Don Hansen, ONR: Given the verification then, and I guess I have to give you a 103 word vocabulary, what do you see as the minimum accuracy that you would accept with a "go" system to go to your agency and say this is it? You've got 997.

A: One of the things which I will find out, with any luck at all, in the next three or four months will be a direct comparison with the key entry system. I do know this, the language that they have to talk with their fingers is sufficiently artificial that a significant fraction of messages which are attempted to be entered with the fingers are flat out rejected because the format is wrong. Some kinds of errors are detectable. If you're controlling a high

altitude sector and you try to put in an assigned altitude of 5,000, it will light up and say "tilt". On the other hand, if you intend (in a high altitude sector) to put in 37,500 and you put in 37,000, this is undetectable. You've got to detect that yourself. So the answer to your question is I honestly don't know. We worked with a guy some years ago who did an ops-analysis type study on the possibility of a mid-air in certain circumstances. He came back with possibly one in a billion operations, or something like that. And his boss said that we will never be able to publish that. The Answer is that we don't have any. We don't intend to ever have any and, in point of fact, that is the truth.

Q: Michael Nye: You quoted a rate, a data entry rate, a manual entry rate where you suggested that in the speech recognition test that you ran, you simulated 70,000 key strokes.

A: That was simply a total aggregate, a key-by-key comparison between 23,000 voice entries which were translated into the equivalent of 69,000 keystrikes.

Q: Michael Nye: I understand. I guess my question is one of two parts. The first part is that we make a big to do about the accuracy of the speech recognition device but we don't talk about the accuracy of the human's ability to be able to enter 70,000 key strokes in the right kind of format.

A: You're absolutely right there. The only people who can do anything like that are professional keypunchers who do nothing but punch keys. Air traffic controllers are never going to be in that business while I'm alive.

Q: Michael Nye: O.K. It's then safe to say that the accuracy, no matter what the accuracy is, as long as its above 98%, is probably more accurate than the manual method.

A: I intend to find that out in the next couple of months.

Q: Michael Nye: The other question was: you made the statement that it does not appear that speech recognition offers any advantage in terms of input speed or throughput and I challenge that and I'm curious.

A: Well, at present if you take a hunt and peck operation where the shifts of attention are involved as well as remembering and constructing this artificial language, we're talking about three or four keys a second. In other words, the coded key equivalent of the message, we're talking about three or four keys a second, that's about what you get out of voice entry on all the preliminary

data I have now. Speed is not so much the name of the game. I can conceive of applications where the speed would be much greater. That just happened to be an observation in passing.

Q: Michael Nye: O.K. I guess the point of view that I had was that you're looking at an isolated application where the speech recognizer is reduced to working in an environment that simulates a keyboard entry routine whereas the real benefit of this kind of technology is in applications where you eliminate the manual method of data capture and you use a voice method. In essence, instead of saying a series of code numbers or code words you actually say the phrase, the phrase is entered in a split second whereas to reduce that to a manual method, maybe four or five key strokes, and in that application, there is a tremendous increase in throughput. Is that true?

A: As I said in the beginning, at least initially I'm working with "unnatural" languages, if you'll pardon the expression.

Q: Don Hanson, ONR: Your vocabulary of 103 may be picked to give a good result. What if you elected to increase the vocabulary? What if you jumped to 1,000, what would you expect your accuracy to be?

A: They probably wouldn't ever do that in the whole class of applications that I'm talking about. The total language of air traffic control, the total human language, is not over 300 or 400 utterances. That's one point. Secondly, as long as your subsets are small enough (and other people have alluded to this) accuracy need not suffer. If you get a subset over 40 or 50 elements and especially, for instance, in this case of fixes, place names, and this sort of thing in the real world you're going to run into things which are sound alikes and you're going to be stuck with them. You're going to get some reduction in accuracy, no doubt about that. Big enough subsets big enough possible set of confusions, errors go up.

Q: Don Hanson, ONR: No, I was thinking of your experience. What's a group figure? 80%, 90.

A: Wouldn't even guess it at a thousand words. No. Not from my experience.

Q: Danny Cohen, Information Sciences Institute, USC: I really appreciate the comment about phrase recognition and I think that the right way to go is by recognizing phrases and being more one mutual language. Does anyone have statistics for recognition of phrases? Is the number still 99.3 or is it more like 60 or 40?

A: Well, I'm working strictly in isolated utterance recognition. Now an utterance can be a whole phrase. When we get into speech understanding where there is some interpretive and analytic work involved in the understanding of what the constituents and the sense of a phrase are you're getting out of my field and I don't know the answer.

Q: Leon Ferber: I wanted to ask you whether you did any studies which probably has nothing to do with speech recognition but there is this phenomenon that either you look at the display or you look at a display and you talk you're just as effective as if you just look or just talk. It's this effect of walking and chewing gum.

A: Yes. We have a little data on this. We've done some work for instance with head mounted eye cameras and this sort of thing and many times the human operator seems to be functioning in parallel fashion or simultaneous fashion but he is really switching back and forth between a couple of sequential operations. I think the truly vital thing is not to lose the picture and if you must look away, you will lose the picture and you will lose some of the important parts of the picture and these are the things that are now possible through computers such as the conflict avoidance alerts. Its one thing to hear a bell and look back and see what's blinking and then you try to figure what the heck it is, another thing when you've got your eyes on that display all the time.

Q: Bob Fleming, Naval Ocean Systems Center: I was wondering if this was introduced to air traffic controllers. Have you given any thought to the mechanics of switching in and out of when he is talking to another aircraft versus talking to the system itself?

A: That is one of the very definite operational problems that we're going to have to face. I see it as a detent microphone switch or something of that sort at least for a starter. One of the things I see in the end going toward the sublime, I see much of what a controller has got to tell the computer with his fingers he has already told a pilot with his mouth. Now if we can just pick out what the computer needs to know from what he told the pilot we have it made. Thank you very much.