LA-UR -81-2052

TITLE: APPLICATION OF COMPUTER VOICE INPUT/OUTPUT

AUTHOR(S): W. Ford
D. G. Shirk

MASTER

SUBMITTED TO: 22nd Annual Meeting of the Institute of Nuclear Materials Management, July 13-15, 1981, San Francisco, California

University of California

LOS ALAMOS SCIENTIFIC LABORATORY
Post Office Box 1663 Los Alamos, New Mexico 87545
An Affirmative Action/Equal Opportunity Employer

APPLICATION OF COMPUTER VOICE INPUT/OUTPUT*

W. Ford and D. G. Shirk
Safeguards Systems Group
Los Alamos National Laboratory
Los Alamos, New Mexico

ABSTRACT

The advent of microprocessors and other large-scale integration (LSI) circuits is making voice input and output for computers and instruments practical; specialized LSI chips for speech processing are appearing on the market. Voice can be used to input data or to issue instrument commands; this allows the operator to engage in other tasks, move about, and to use standard data entry systems. Voice synthesis [illegible] can generate audible, easily understood instructions. Using voice characteristics, a control system can verify speaker identity for security purposes. Two simple voice [illegible] systems have been designed at Los Alamos for nuclear safeguards applications. Each can easily be expanded as time allows. The first system is for instrument control that accepts voice commands and issues audible operator prompts. The second system is for access control. The speaker's voice is used to verify his identity and to activate external devices.

I. INTRODUCTION

Voice recognition and voice synthesis are now sufficiently developed for use in varied applications. The use of voice provides several advantages over other man-machine communications methods. It is a completely natural, straightforward technique. No special operator skills are required. Hands and eyes are free to do other tasks and freedom of movement is allowed. The operator is not required to carry special devices.

Two prototype systems were developed at Los Alamos for use in safeguards systems. These two systems were selected as representative of a large range of future uses. The development was intended to prove feasibility and reliability; specific applications would require individual study and system customization.

Commercial voice recognition/synthesis hardware is now available from various manufacturers. The host computer required for voice input/output can range from a microprocessor to a standard minicomputer or large scale system. Applications software must be written for the specific application. High-level languages can be used for most applications.

The two specific applications of voice recognition that were examined are (1) data entry/instrument control and (2) speaker voice verification. For data entry or instrument control, the system requires [illegible] recognition and few rejections of valid words. Some latitude in voice variations in order to [illegible] speakers is desirable. For speaker verification or other security applications, the system would accept a higher rejection rate of valid speakers to minimize acceptance of [illegible].

II. BASIC VOICE RECOGNITION [illegible]

At the present time available systems are for isolated-word, speaker-dependent use; that is, the user's voice must be enrolled and stored in the system. Speaker independence can be obtained if the vocabularies are very limited and all users speak the same dialect. The number of bits used for recognition is considerably [illegible] from that for synthesis. Only a few key parameters are used and are reduced to binary form. The selected parameters are sampled at various times during the speech interval. For example, an utterance of from 0.1 to 1.2 seconds [illegible] only 128 data characters of 4 bits each. This could be 16 parameters sampled every few milliseconds. The entire duration is divided into 8 intervals, and the average of all samples during the interval is calculated. The bit patterns, called templates, are obtained from an average of the utterance spoken several times by the user.

For recognition, an unknown utterance is processed in the same manner as that to obtain the templates, and the pattern is compared with all templates. Scores are computed indicating the degree of match between the unknown and the templates. The template having the highest score is noted. If the match exceeds some threshold value (adjustable in the system), the unknown is considered to be the utterance that produced the template. The template having the second highest

*This work was supported by the US Department of Energy/Office of Safeguards and Security.

score can also be noted and be required to differ from the highest score by a given amount (also selectable). This we refer to as a confidence level.

III. HARDWARE DESCRIPTION

Hardware was chosen that was commercially available, easy to use, low cost, and compact. A small amount of computing power is needed, which can be supplied by a microprocessor. If voice is added to an existing system, the existing computer can serve as host with very little overhead. The same set of voice hardware and host computer was used for both systems. For voice recognition we are using an Interstate Electronics VRM-102 with an RS-232C serial interface. The recognition vocabulary is 100 words. The microphone is a head-worn, close-talking, noise canceling Shure model SM-10. The voice response function is provided by two circuit boards from the Texas Instruments TMS9900 microprocessor line: a /306 Speech Module and a /101 CPU Module. The Speech Module has a fixed 177-word vocabulary. CPU software to drive the Speech Module is written in assembly language and contained in EPROM. Communications are through an RS-232C serial interface. The host computer is a Digital Equipment LSI-11/2 microprocessor with dual floppy disk. Most of the software was written in FORTRAN. Asynchronous device drivers were written in assembly language.

IV. VOICE-CONTROLLED INSTRUMENT

The first application is a voice-controlled instrument. A block diagram is shown in Fig. 1. At instrument turn-on, the operator is requested, by the voice synthesizer, to enter his operator ID number. The ID is spoken as a string of isolated digits. If a digit is not recognized, the operator is informed of the failure both on a terminal and by the voice synthesizer. If all the digits are recognized, but the ID number is invalid, the operator is notified in the same manner. Acceptance of a valid ID activates the instrument. At the completion of the instrument task, a voice response alerts the operator.

Some easily added extensions to the voice controlled instrument system were considered but not implemented. Included in these extensions are (1) a multiple-user environment in which the user's voice templates are loaded into the recognition module upon his acceptance as a valid user and (2) an expanded command list to allow several additional instrument operations such as calibrate, analyze, etc. Each operation is initiated by a spoken command.

The recognition vocabulary can be partitioned so that only a portion would be active at any given time. Initially, only the digits needed to enter operator ID would be active. Upon acceptance of the ID, commands for the various instrument functions would be active. These could be CALIBRATE, ASSAY, MEASUREMENT CONTROL, DIAGNOSTIC, HELP, and LOG, for example. If ASSAY were spoken, then a vocabulary consisting of the digits, BACKGROUND, SINGLE PASS, and MULTIPLE PASS would be enabled. Such a strategy along

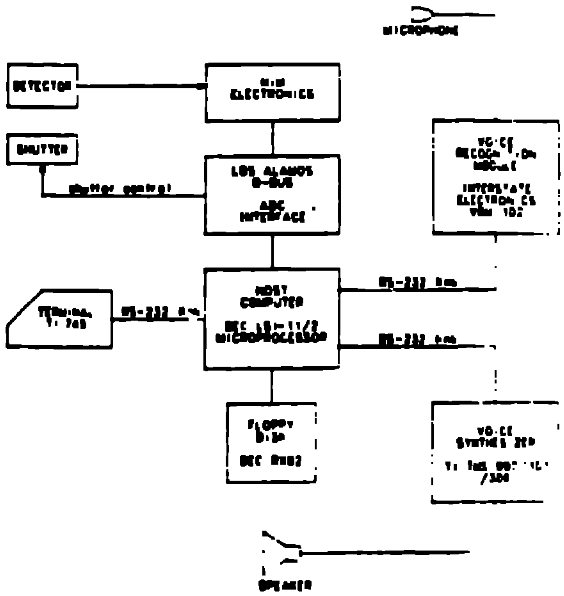

Fig. 1. Voice-controlled instrument block diagram.

with careful vocabulary selections and proper threshold setting can provide good word discrimination, few rejections of valid words, and almost never a misrecognition. The partitioning scheme described above also allows the same word to have two different meanings. For example, under DIAGNOSTIC the word TERMINAL would initiate a diagnostic of the system terminal. Under LOG, TERMINAL would direct the output of the system log to the system terminal. However, the word TERMINAL would have to appear as two separate vocabulary items.

Our experience with the voice actuated-instrument control system over a period of three months with two validated users in the laboratory, indicated that the system was very reliable and efficient. Statistics were not collected on detailed operation. Outright rejection of valid words remained acceptable (less than 1 out of 4), and misrecognition was never a problem. Retraining was performed once. The instrument already contained an LSI-11/2 microprocessor with dual floppy disk, which was used as the host computer. For multiple users, templates can be easily stored on disk. This technique was not used on the instrument but would be identical to the method used in the system to be described next.

V. ACCESS CONTROL SYSTEM

The second application is designed to illustrate voice access control techniques. Each user produces a set of voice templates using the same set of utterances for all users. Originally, the system quiescent state was to contain a template of the same utterance from all users. Upon speaking that utterance, the speaker would be identified and his vocabulary downloaded from disk. Upon completion of verification the quiescent vocabulary would be restored. While this method has merit and may be desired in some applications, the problem of getting an invalid user past the quiescent state made data collection slow and difficult. Therefore, that scheme was replaced with typing in an ID number, unique to the user, on a terminal. This action simulates a badge reader or other identification means. The voice then verifies the identity, for example, to prevent use of a stolen or forged badge.

Acceptance of the ID loads the user templates from disk to the recognition unit. This loading process takes 2 to 3 seconds. The user's vocabulary consists of 16 words that were selected from those available on the voice synthesizer and such that the vowel sounds would be distinct. It is in the vowel sounds that the differences in speech between individuals is most prominent. The word list is given in Table I. One of the vocabulary items is randomly selected, and the user is prompted via the voice synthesizer to speak the word. On an entry try, the user must have three vocabulary items recognized to gain access. Nonrecognition of four vocabulary items results in denial of access. The technique is illustrated by movement in a 3 by 4 unit space with absorbing boundaries as shown in Fig. 2. A right or up unit movement occurs at each node as the result of a word recognition attempt. The object is to reach the right side (acceptance) before reaching the top (rejection.) Thus the user is prompted to speak from 3 to 6 vocabulary items. Three recognitions with 0 to 3 nonrecognitions grant access and four nonrecognitions with 0 to 2 recognitions deny access. This process takes from 3 to 7 seconds, depending on the number of recognitions and the user reaction time. The user is not informed of the results of any individual vocabulary item. He is informed via the voice synthesizer of the overall result. Granting access means performing any task within the capabilities of the host computer and its peripherals. The access information could be used in other ways such as personnel inventory if egress were also controlled. Slightly over 1 kilobyte of storage is required for each user's templates. Thus, the number of users depends on the storage space available for templates.

TABLE I

VOCABULARY FOR VOICE VERIFICATION

Word No.	Utterance	Word No.	Utterance
0	CALL	8	OFF
1	HERTZ	9	POINT
2	HIGH	10	POUND
3	HUNDRED	11	TOOL
4	MOVE	12	REPEAT
5	MANUAL	13	START
6	MEGA	14	WAIT
7	NORTH	15	ZERO

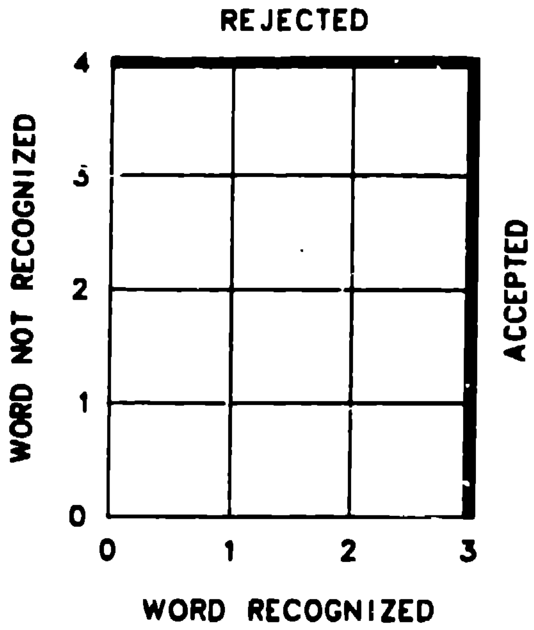

Fig. 2. User acceptance model.

In a speaker verification system there are two types of errors of interest: Type I, rejecting a valid speaker, and Type II, accepting an invalid speaker. There is some tradeoff between the types of errors. It is possible to change system parameters dynamically to balance errors for the particular current task.

VI. RESULTS OF ACCESS CONTROL SYSTEM

The prototype access control system is still under development and evaluation. It is being used to evaluate strategies, vocabularies, and equipment. One configuration was selected for extensive testing and data collection. This configuration is shown in Fig. 3 using the vocabulary of Table I. Access occurs upon 3 recognitions before 4 rejections or denial upon 4 rejections before 3 recognitions. The acceptance threshold was set at 117 with a perfect match being 128, a confidence level greater than 10, and a check of recognized word versus prompted word. Seven valid users (6 male and 1 female) trained the system, i.e., produced vocabulary templates. Sixteen different individuals, 13 males and 3 females, were used as imposters trying to gain access through the use of a valid ID. The data are presented in Table. II. The imposter success rate was 2.8%. Refinements to the system can reduce this rate by an order of magnitude. As shown in Table III, 10 of the 25 successes by imposters involved the repeat of a high-probability word. If the software were modified to eliminate repeats on any given trial, success rate would drop to 1.7%

Analysis of the data on a word-by-word basis shows where improvements in vocabulary selection can be made. Recognition success for each individual word is tabulated in Table IV, in order of increasing recognition success. Data in Table III show that only two or three certain

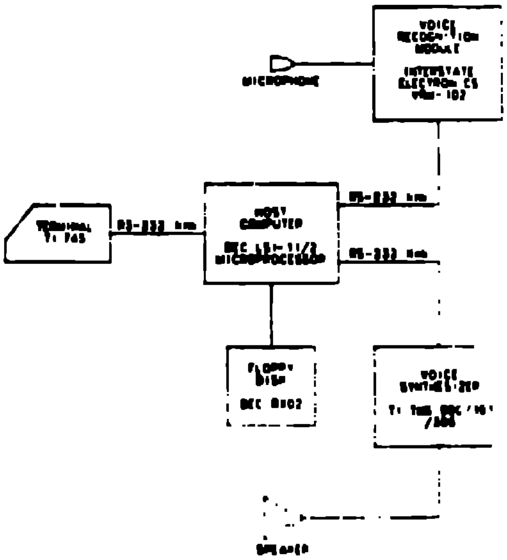

Fig. 3. Voice-controlled access system block diagram.

TABLE II

IMPOSTOR SUCCESS (EXPERIMENTAL)

Impostor Number	Male or Female	Number of Tries	Number of Successes
1	F	[illegible]	0
2	F	60	1
3	M	62	2
4	M	56	0
5	M	60	1
6	F	60	0
7	M	41	1
8	M	61	5
9	M	60	1
10	M	20	0
11	M	31	4
12	M	6[illegible]	8
13	M	40	0
14	M	43	2
15	M	120	0
16	M	49	0
		846	25

TABLE III

WORDS RECOGNIZED IN IMPOSTOR SUCCESSES

Imposter Number	Against Valid User No.	Words Recognized
9	4	CALL, HERTZ, HERTZ
11	2	HIGH, CALL, START
11	2	TOOL, START, REPEAT
11	2	MANUAL, TOOL, ZERO
11	2	REPEAT, HERTZ, TOOL
12	2	HUNDRED, ZERO, HERTZ
12	2	ZERO, HUNDRED, CALL
12	2	CALL, NORTH, MEGA
12	3	MEGA, ZERO, MEGA
12	5	START, REPEAT, REPEAT
12	7	HUNDRED, HIGH, ZERO
12	7	MANUAL, HERTZ, HUNDRED
12	7	START, REPEAT, HUNDRED
14	2	MEGA, START, ZERO
14	3	CALL, CALL, HERTZ
2	4	HIGH, ZERO, WAIT
3	3	HERTZ, REPEAT, REPEAT
3	3	HERTZ, REPEAT, REPEAT
5	6	POINT, HERTZ, HERTZ
7	2	HERTZ, HUNDRED, WAIT
8	2	ZERO, HERTZ, HERTZ
8	3	MEGA, MEGA, REPEAT
8	7	HUNDRED, REPEAT, HERTZ
8	7	REPEAT, REPEAT, HUNDRED
8	7	MOVE, ZERO, REPEAT

TABLE IV

WORD RECOGNITION PROBABILITIES (EXPERIMENTAL)

No.	Word Number	Word	Number of Times Prompted	Number of Times Recognized	Recognition Probability	Cumulative Probability
1	[illegible]	[illegible]	[illegible]	0	0	0
2	[illegible]	NORTH	[illegible]	4	[illegible]	[illegible]
3	8	OFF	[illegible]	5	[illegible]	[illegible]
4	11	TOOL	[illegible]	6	[illegible]	[illegible]
5	[illegible]	WAIT	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	MOVE	[illegible]	[illegible]	[illegible]	[illegible]
7	[illegible]	POINT	[illegible]	[illegible]	[illegible]	[illegible]
8	5	MANUAL	[illegible]	11	[illegible]	[illegible]
9	[illegible]	START	[illegible]	[illegible]	[illegible]	[illegible]
10	[illegible]	MEGA	[illegible]	[illegible]	[illegible]	[illegible]
11	[illegible]	HIGH	[illegible]	[illegible]	[illegible]	[illegible]
12	[illegible]	HUNDRED	[illegible]	[illegible]	[illegible]	[illegible]
13	[illegible]	ZERO	[illegible]	[illegible]	[illegible]	[illegible]
14	[illegible]	REPEAT	[illegible]	[illegible]	[illegible]	[illegible]
15	[illegible]	CALL	[illegible]	[illegible]	[illegible]	[illegible]
16	1	HERTZ	[illegible]	[illegible]	[illegible]	[illegible]

words had a high probability of being recognized in each individual case, different words for different pairings of impostor versus valid user. The bar graph of Fig. 4 shows the number of times a vocabulary word was involved in an impostor success. Revision of the vocabulary will result in further improvement. In every case an impostor gained access, one or more of the four highest-probability words was recognized. Thus, eliminating these four words should drastically reduce the impostor success rate.

To estimate impostor-success probability as a function of individual word-recognition probability, a theoretical study is underway and a Monte Carlo computer simulation is being designed. In the meantime, some qualitative results can be obtained by making several simplifying assumptions and treating the process as a stochastic process. The model differs from the experimental data because word recognition is a deterministic process, and the results are binary events. The study is useful, however, for providing insight into possible improvements.

It is assumed that all 16 words have an equal probability, p, of being recognized. Requiring 3 recognitions before 4 failures gives 20 possible sequences: 1 sequence of 3 words, 3 sequences of 4 words, 6 sequences of 5 words, and 10 sequences of 6 words. Therefore, the probability of successfully gaining access is given by:

$$P = p^3 + 3p^3(1-p) + 6p^3(1-p)^2 + 10p^3(1-p)^3 \qquad (1)$$

A plot of this function for small values of p is shown in Fig. 5. We can use the data of Table IV to calculate experimental values for p and P. Values for p were calculated from

$$p = \bar{p}_n = \frac{1}{n} \sum p_i \qquad (2)$$

where $\bar{p}_n$ = average recognition rate for first n words of Table IV.

n = number of words averaged

p_i = recognition rate for word i.

P_n was calculated by using only those impostor successes containing the first n words of Table IV. Four points are plotted in Fig. 5:

n	$\bar{p}_n$	P_n
16	.0775	.0283
15	.0683	.0158
14	.0625	.0124
13	.0559	.0057

Experimentally, P was zero for n less than or equal to 12. A P of zero is probably not obtainable, but extrapolation of experimental data on Fig. 5 gives P < 0.0025 for n = 12. (A quadratic least-squares fit of the form $P = a_0 + a_1 p + a_2 p^2$ was found, and P was calculated for p = 0.04929.)

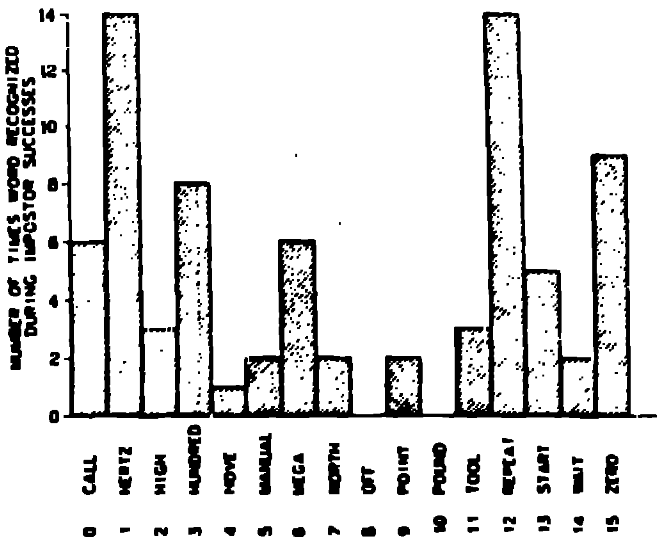

Fig. 4. Total recognitions during imposter success.

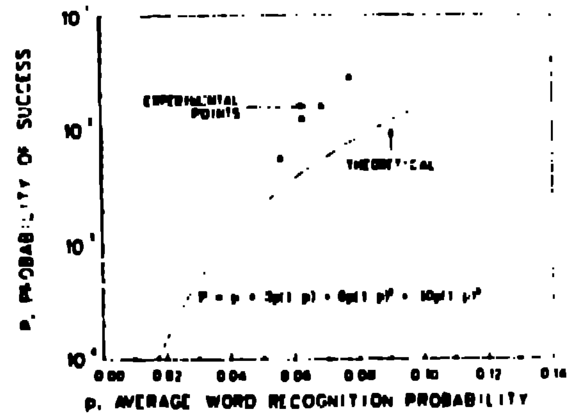

Fig. 5. Probability of success.

The experimental results are also shown in the bar graph of Fig. 6 in which impostor success and number of trials against each valid user is shown. All others had zero success. Fig. 7 presents a bar graph showing total successes against each valid user. The number at the top gives the total number of trials. An intruder scenario might be much different than the tests. First, the intruder would have to steal or otherwise obtain a valid badge (assuming a badge reader is used). Then he would have to achieve success within the number of trials allowed. A failure could automatically alert a guard or other operator.

Another consideration is the success or failure of a valid user. Continued rejection of a valid user would not be acceptable. Our experience shows that a valid user requires an average of 4 to 4.5 utterances to be accepted. From equation (1), a p of only 0.8 gives a P of 0.983. Occasionally however, a valid user is

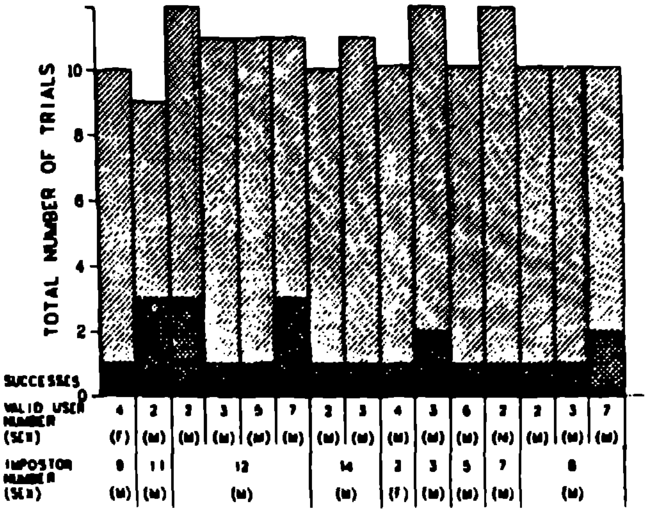

Fig. 6. Impostor success versus number of trials for each valid user.

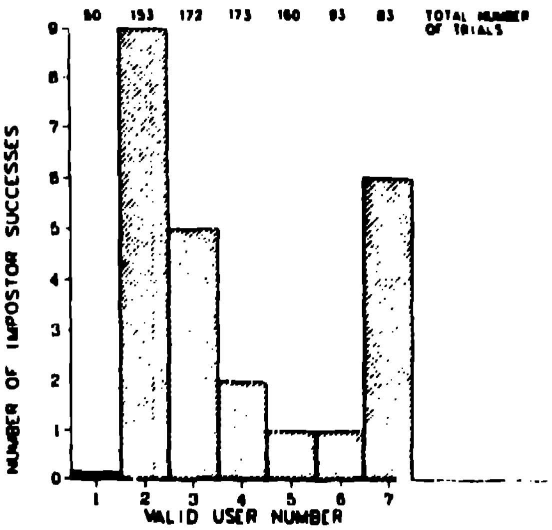

Fig. 7. Number of successes against each valid user.

rejected. This rate goes down as familiarity with the system is gained and the users learn to speak consistently. A cold or allergies can produce enough change to cause trouble, and provisions must be made to accommodate these temporary cases. Long-term voice changes are accommodated by the use of periodic retraining. The training process requires about 10 minutes for a first-time user and drops to 5 minutes or less for experienced users. The process of repeating a 16-word vocabulary 5 times actually consumes about 2 minutes. Prompting the user for both training and in actual use with the voice synthesizer helps stabilize pronunciation and encourage consistency of speech, especially if words tend to be pronounced in more than one way. For example, the synthesizer pronounces HERTZ as "hurts" rather than "hairtz." Nervousness, frustration, or other types of stress definitely reduce a valid user's success. Voice recognition is not for stressful situations!

VII. CONCLUSIONS

There has been tremendous progress in the field of voice synthesis and speech recognition in the past few years. We have shown that speech technology using state-of-the-art equipment can be a very effective tool in safeguards and many other systems.

Careful selection of vocabulary and a proper control strategy produced a successful access control system. Data entry and instrument control were readily accomplished.

VIII. ACKNOWLEDGEMENT

The authors are indebted to Aaron Goldman for helpful discussions concerning the statistical analysis, and to our colleagues who volunteered their voices for the tests.

REFERENCES

1. N. Rex Dixon, and Thomas B. Martin, Automatic Speech & Speaker Recognition, IEEE Press, New York, 1979.

2. George R. Doddington, "Personal Identity Verification Using Voice," Proceedings of ELECTRO-76, May 11-14, 1976, pp. 22-4, 1-5.

3. R. W. King and T. D. Barnes, "Advanced Access Control System," Proceedings of 21st Annual INMM Meeting, Palm Beach, Florida, June 30-July 2, 1980, pp. 362-9.