THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 59, No. 7, September 1980
Printed in U.S.A.

A Voice-Controlled, Repertory-Dialer System

By L. R. RABINER, J. G. WILPON, and A. E. ROSENBERG

(Manuscript received January 17, 1980)

This paper describes a speaker-trained, voice-controlled, repertory-dialer system that includes a real-time speech analyzer, an isolated-word recognizer, a voice-response system, and a simulated dialer. The system is implemented on a minicomputer with a high-speed array processor performing the real-time operations. The vocabulary consists of seven command words, ten digits, and any number of names up to a specified maximum. To train the system, the user speaks each vocabulary word twice to provide reference templates for the system. After training, the system can dial the telephone number corresponding to any name in the repertory, or it can dial a 4-digit telephone extension spoken as a string of isolated digits. The system operates in two modes. In the first (cued by a double beep), the user can modify the repertory either by adding or deleting names or by changing a phone number, or the user can enter the second mode using a specified command word. In the second mode (cued by a single beep), the user can speak any name in the repertory or can give a 4-digit telephone extension. The system was tested by six talkers (three male and three female, three of whom were naive and three experienced users) over a three-week period. A total of 4620 words were spoken, and during the course of the text there were no recognition errors. A request for a repeat of a spoken word occurred about 2 percent of the time.

I. INTRODUCTION

Research in isolated word recognition has progressed to the point where it is now feasible to implement simple, but useful, task-oriented recognition systems. It is the purpose of this paper to describe one such system, a voice-controlled repertory dialer, that has been implemented in the Acoustics Research Department at Bell Laboratories.

Before describing the operation of the repertory dialer, it is worthwhile reviewing the state of the art in isolated word recognition. Word-

recognition systems may be classified according to the following factors (among others):

(*i*) Whether speaker-trained or speaker-independent.

(*ii*) Size and complexity of vocabulary.

(*iii*) Operating environment.

(*iv*) User training required to use the system.

The most accurate and reliable recognizers are those that are speaker-trained, with small to moderate size vocabularies (10 to 50 words) of low complexity (i.e., all words are distinctly different), with a high quality recording system, a low background noise environment, and a modest amount of user training. Such recognizers can reliably maintain accuracies of 99 percent or better across a wide range of users.[1] As some factors are degraded (e.g., larger vocabularies, more complexity, telephone inputs), the reliability and accuracy of the recognizer tends to become worse. To maintain the high reliability and accuracy required for a practical system, even when the most ideal set of recognition factors cannot be obtained, the context of the recognition task must be relied on to detect and correct recognition errors or to provide user feedback for a repetition of the voice command when an otherwise unreliable recognition would be made. An example of such a task-oriented recognizer is the directory assistance system proposed by Rosenberg and Schmidt.[2] In this system, in which a user requests directory information by spelling the person's name (letter by letter), the recognition accuracy on the letters is only about 70 to 80 percent; however, for an 18,000-name directory, the name accuracy is close to 98 percent.[3]

The ways in which the task-oriented recognizer can improve the accuracy and reliability of the word recognizer are:

(*i*) The use of partitioned vocabularies. At each step in the task, the word to be recognized falls into a subset of the entire recognition vocabulary; hence, only this subset need be searched for the word. Using this technique, the effective vocabulary size and complexity can often be substantially reduced.[4]

(*ii*) The use of semantic constraints in the task to correct errors in recognition. For example, if a time of the day is requested (a 2-digit sequence) and the sequence 37 is recognized, the task knows such an hour is impossible and can find the most likely candidate consistent with the semantic constraints of a time of the day.

(*iii*) The use of a rejection threshold on the distance score in which no recognition candidate is accepted, causing a request to the user to repeat the command. In this manner, the recognizer, and/or the task, can detect cases in which reliable recognition is in doubt (either because the recognition scores are poor, or because it is impossible to decide between two or more candidate words) and, rather than make an unreliable decision, it can pass the burden back to the user.

The system to be described in this paper, a voice-controlled repertory dialer, makes use of all the above techniques to provide a very accurate and reliable recognition system.

Section II describes the basic operation of the dialer and Section III describes a series of tests that measured the performance of the dialer in a realistic operating environment. In Section IV, we provide a brief discussion of the results and summarize the main contributions of the work.

II. OPERATION OF THE REPERTORY DIALER SYSTEM

Figure 1 is a block diagram of the repertory dialer system. The main elements of the system are:

(*i*) A real-time speech analyzer that detects the presence of speech on the input line and analyzes the speech to give features [frames of LPC (linear predictive coding) parameters] appropriate for the word recognizer.

(*ii*) An isolated word recognizer (of the type originally proposed by Itakura[5] that compares the spoken word to a subset of the words in a template store and provides an ordered list of word candidates.

(*iii*) A voice response system to provide spoken commands to the user to guide the use of the repertory dialer system. The voice response system has both read-only memory (ROM) storage (for prerecorded words and phrases) and random access memory (RAM) storage (for the names in the directory).

(*iv*) A word template store for storage of the reference patterns for each word of the vocabulary.

(*v*) A directory store containing the current set of repertory names and the associated telephone numbers.

(*vi*) A dialer to outpulse the desired telephone number.

(*vii*) Logic control to:

- (*a*) Guide the voice response system.
- (*b*) Provide auditory commands and feedback to the user.
- (*c*) Guide the word recognizer in deciding which subset of words is required for recognition.
- (*d*) Dial a telephone number when required.
- (*e*) Control storage in the template store (in the training mode) and the directory store (when adding or deleting names, or modifying telephone numbers).

(*viii*) A mode switch to set the system for training (i.e., creation of word reference templates) or testing (normal usage mode after training).

2.1 Training the system

To use the system, it must first be trained. To do this, the mode switch is set to 0 (training), and the voice response system prompts

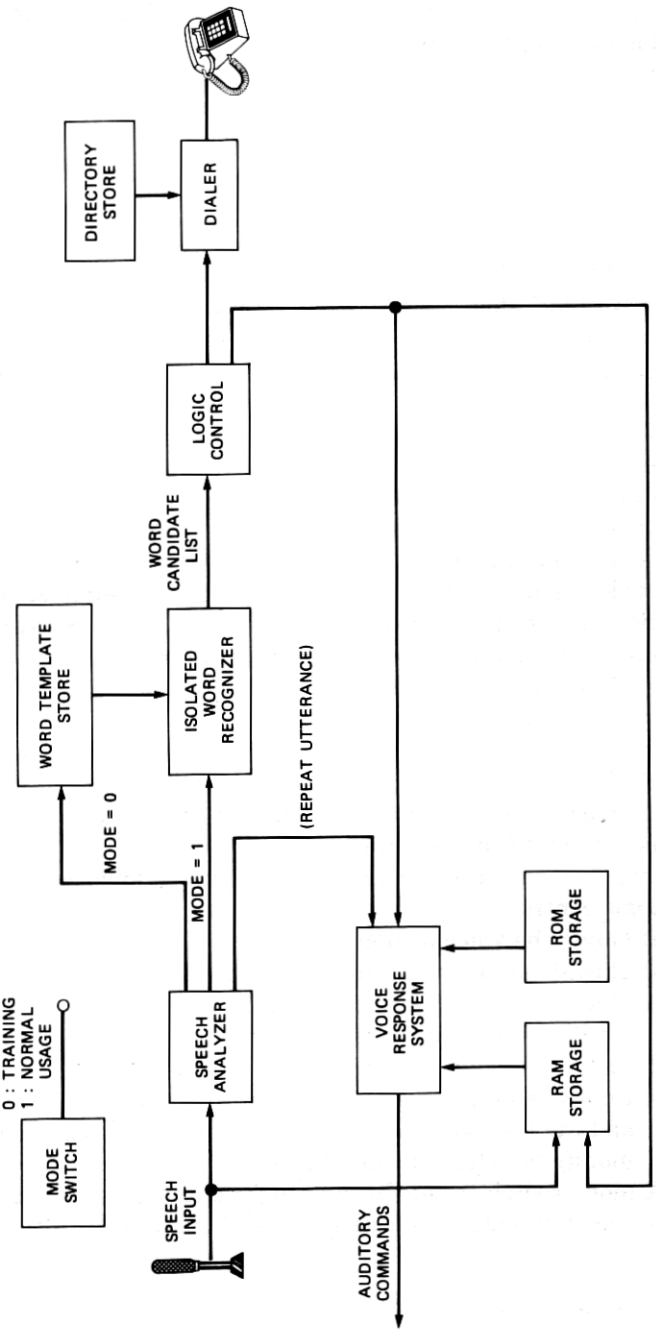

Fig. 1—Block diagram of the voice-controlled repertory-dialer system.

the user to say each word of the vocabulary at a given auditory command (a beep). Table I lists the vocabulary used to evaluate the system. It consists of 7 command words, 10 digits, and 20 names of people at Bell Laboratories. Each training word is analyzed to give a set of LPC features for each frame (45-ms frames with a 30-ms frame overlap between adjacent frames) in the word, and these features are stored (as sets of autocorrelated LPC coefficients) in the word template store. If the speech analysis box detects any recording problems (e.g., level too low, no word spoken, artifacts in the recording), the user is requested to repeat the word. Complete training of the system consists of two replications of all the vocabulary words.

2.2 Normal use of the system

Following training, the system can be used as a voice dialer of any telephone extension (a 4-digit code spoken as a string of isolated digits), or as a repertory dialer for the names entered in the training mode. It can also be used to add to or delete names from the repertory and to modify the telephone number of a name in the repertory.

To use the system, the mode switch is set to normal usage (MODE = 1), and the voice response system cues the user to speak a command word by sending a double beep. The user has up to 18 seconds to speak one of the command words as an isolated word. If, after 18 seconds, no isolated command word is found, a double beep is again sent to the user and the process is repeated until one of the command words is recognized.

The set of command words, and the action taken is as follows:

OFFHOOK: Take the telephone off hook preparatory to dialing a

Table I—Words used in the repertory dialer vocabulary

OFFHOOK		ALLEN	
HANGUP		BAKER	
MODIFY	Command	BERKLEY	
DELETE	Words	COKER	
ADD		CROCHIERE	
ERROR		FLANAGAN	
STOP		HALL	
0		HANNAY	
1		JAYANT	
2		LEVINSON	Names
3		MATHEWS	
4	Digits	MCGONEGAL	
5		MOORE	
6		PRIM	
7		RABINER	
8		ROSENBERG	
9		SONDHI	
		UMEDA	
		WEST	
		WILPON	

number. System responds with a single beep to prompt the user to speak a name in the repertory or a string of four isolated digits.

HANGUP: Terminate use of the system; hang up the telephone.

MODIFY (NAME): Change the telephone number of repertory name (NAME). System guides user (via voice response commands) to speak new telephone number.

DELETE (NAME): Delete entry (NAME) from directory and RAM storage (in which the spoken form of the name is stored).

ADD: Add a new name to the repertory. System first requests name to be added and stores the speech waveform (in coded form) in the RAM storage of the voice response system. The system next requests two replications of the name for the word template store. Finally, the system requests the telephone number (a 4-digit extension) for the directory store.

ERROR: The system disregards the most recently recognized word (for which an error occurred—on the part of either the user or the recognizer), and the user is requested to repeat the actual word. The ERROR command can be used after any recognition made by the system, because the system verifies each recognition via the voice response system and follows the response with a cueing beep. The user can say ERROR after the beep occurs.

STOP: The system goes back to the command mode and disregards the current command.

The command words MODIFY, DELETE, and ADD operate off line (i.e., the telephone is on hook) and affect the directory, the template store, and RAM voice response storage. At each stage in the execution of these commands, user verification of the recognition input is requested. For ease of use, a correct recognition is verified by silence; the ERROR command is used in the case of an error.

The command word OFFHOOK puts the system in the dialing mode. After the single-beep cue, the user can speak a name in the repertory or a string of 4 isolated digits.

2.3 Features of the repertory-dialer system

Several key points should be made about the voice dialer. The first is that all communication between the user and the system is by voice. No visual display of any type is needed to train or to use the system. The voice response commands (as stored in ROM memory) include the 13 phrases shown in Table II and the 7 command words and 10 digits of Table I. RAM memory space has to be allocated for each of the names in the repertory—typically, 10 to 20 names should be sufficient. If all voice response commands are coded to 24 kbps,[6,7] using a waveform coding technique like ADPCM, a total of about 720,000 bits (30 seconds × 24,000 bps) are required for ROM storage, and 360,000

Table II—Phrases used by the voice response system

1. After each tone say the specified word.
2. Please repeat.
3. Please repeat the command.
4. Please repeat the number.
5. At the beep, speak the name to be added.
6. Please repeat the name to be added.
7. At the beep say the word (—).
8. Please enter phone number.
9. Please repeat the name to be deleted.
10. Please enter the name to be deleted.
11. Please enter new phone number.
12. Please verify.
13. Please repeat the name whose phone number is to be changed.
14. "Beep"

bits (15 seconds × 24,000 bps) are required for RAM storage. With LPC coding, the storage requirements are reduced further by a factor of 10 or more (although the coder-decoder costs increase substantially).

A second feature of the dialer is that the system responds only to isolated word inputs. Thus the user may hold a conversation while the dialer is operating, and the system will not be triggered unless an isolated version of one of the command words is recognized. As mentioned earlier, in order for a word to be recognized, it must have a distance score within prescribed limits and it must have a considerably smaller distance than the next likely recognition candidate. The likelihood of such events occurring during conversational speech is very small.

Another aspect of this system, also mentioned previously, is that the vocabulary of Table I is partitioned for recognition into the following sets:

(*i*) SET 1–7 command words.
(*ii*) SET 2 –20 names, 10 digits, word STOP.
(*iii*) SET 3–10 digits.
(*iv*) SET 4–STOP and ERROR.

Thus, in the worst case, the recognizer must choose among 31 possible candidates. However, even for that case, more information is present in the task. If the recognizer finds a digit, the task knows that it must be part of a 4-digit string. If no such string is found, the task can choose the best recognition candidate among the set of the names and the word STOP. Similarly, if a string of digits is spoken and the recognizer matches the first digit to a name (e.g., 4 becomes MOORE), the task can correct the word to the most likely digit based on the recognition of three subsequent digits.

Finally, it should be noted that the voice repertory dialer system is suited to a wide variety of input devices (telephone, microphone, wireless microphone) and operating environments. It has been informally tested in both large and small rooms (offices and conference

rooms) and formally tested in a computer-room environment. In the next section, we describe the formal test of the system.

III. TESTING THE REPERTORY DIALER

The voice-controlled repertory dialer of Fig. 1 was implemented on a laboratory computer (a Data General Eclipse Computer) using a high-speed array processor (the CSP MAP 200) to perform the real-time analysis and the recognition distance calculations. A wireless microphone was used at the input to simulate a cordless telephone that might be used in an office environment. Hence, the user was not required to be in close proximity to the computer.

The vocabulary of Table I was used as the training set (including the 20 specified names). Six subjects were used to test the dialer. Three subjects were male, three were female. Three subjects were experienced users of speech recognition systems (although not this particular system), and three subjects were naive users. No remuneration was given to the subjects, although all could be considered cooperative users.

The tests were carried out in a computer-room environment. Each subject trained the system and then participated in a performance test that lasted from 2 to 4 weeks, depending on the availability of the subjects. Table III shows the series of commands used by each subject to test the dialer. Each subject executed the commands in sequence once per session for 10 sessions. Each test nominally consisted of 17 full commands with a total of 77 words per test. If errors were made or repeats were requested, the number of words per test increased.

Table III—Summary of commands used to test the repertory dialer system

1. (DB) OFFHOOK — (SB) BAKER
2. (DB) OFFHOOK — (SB) FLANAGAN — (SB) ERROR — (SB) ROSENBERG
3. (DB) OFFHOOK — (SB) 2-3-7-9
4. (DB) OFFHOOK — (SB) HANNAY — (SB) ERROR — (SB) WILPON
5. (DB) OFFHOOK — (SB) 6-0-1-4 — (SB) ERROR — (SB) 1-2-7-4
6. (DB) OFFHOOK — (SB) RABINER — (SB) ERROR — (SB) ALLEN
7. (DB) ADD — (SB) "GRECCO" — "GRECCO," "GRECCO" — (SB) 3-9-4-6
8. (DB) OFFHOOK — (SB) GRECCO
9. (DB) DELETE GRECCO
10. (DB) MODIFY WEST — (SB) 9-5-8-5
11. (DB) OFFHOOK — (SB) WEST
12. (DB) MODIFY HALL — (SB) 8-0-5-2
13. (DB) OFFHOOK — (SB) LEVINSON — (SB) ERROR — (SB) MATTHEWS
14. (DB) OFFHOOK — (SB) STOP
15. (DB) OFFHOOK — (SB) BERKLEY — (SB) ERROR — (SB) UMEDA — (SB) ERROR — (SB) SONDHI — (SB) ERROR — (SB) CROCHIERE
16. (DB) OFFHOOK — (SB) MOORE — (SB) ERROR — (SB) JAYANT — (SB) ERROR — (SB) COKER — (SB) ERROR — (SB) MCGONEGAL — (SB) ERROR — (SB) PRIM
17. (DB) HANGUP

(DB) = > Double Beep
(SB) = > Single Beep

An examination of the material in Table III shows that 30 of the 77 words in the test were command words, 24 words were digits, and 23 words were names. The words OFFHOOK and ERROR (the two most important commands) occurred 12 times each per test. The digits occurred two or three times each per test, and each name occurred at least once per test. During the test, one name was added, one name deleted (the one that was added), and two phone numbers were modified.

3.1 Training results

The training for each subject occurred in the first session and took, on the average, 9 minutes to enter 2 replications of the 37 words. During the training session (which was guided by the voice response system), an average of one request for a repetition of a word occurred during the 9-minute period.

3.2 Test results

During the course of the tests, a nominal total of 4620 words (77 words/test × 10 tests × 6 speakers) were spoken and recognized. However, due to repeated digit strings (when one or more digits were recognized before a problem was detected), an extra 72 words were spoken and recognized during the tests. Of the 4692 recognitions made by the system, no recognition errors were made. The reasons for this high accuracy score have been discussed previously and are emphasized in Fig. 2, which shows plots of the average recognition distance for each speaker for the first recognition candidate (the correct word) and for the second recognition candidate. Also included in each curve are brackets indicating the one standard deviation range (across tests) for each talker. It is readily seen that a large separation exists between the average distance of the first and second candidates for all speakers.

An important question about the test results is how often the system requested a repeat of a word. Each such request averts a potential recognition error. During the course of the test, a total of 106 requests for the repeat of a word occurred. Of these cases, 98 requests for repeats came from the acoustic recognizer. Such cases were primarily due to responses that came before the acoustic cue (the beep), or those that were missed entirely (i.e., the speaker inadvertently said nothing during the recording interval). Although a continuous recording was used, the system would recycle if 2 or more seconds of silence (signal level below a threshold) were detected. Thus, in only 8 cases out of 4700 recordings, the recognizer detected distances that were too large and requested a repetition of a word (or sequence of words). The overall average rate at which a request for a repeat occurred was about 2 percent.

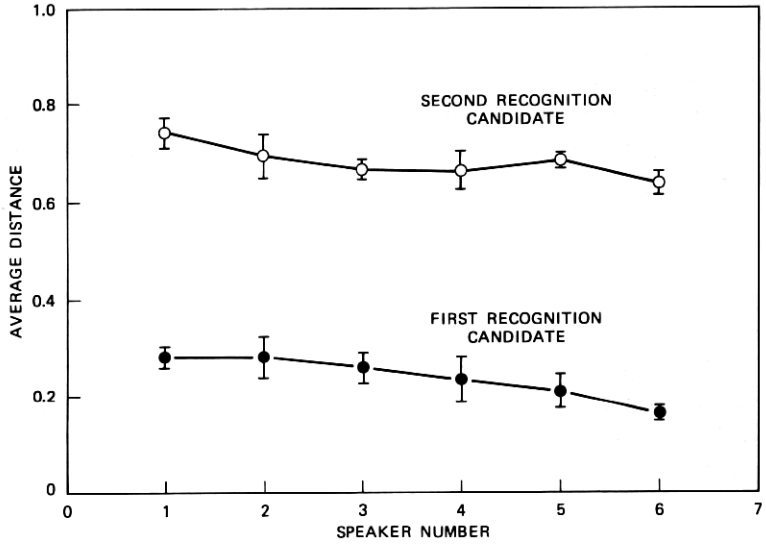

Fig. 2—Average distance as a function of speaker number for the first and second recognition candidates.

The only other statistic that was monitored during the tests was the average time for each test. On the average, a complete test took about 12 minutes, or about 8 to 9 seconds per recognition, prompting, response, and verification. Considering the amount of communication that takes place between the user and the system, such average times seem quite reasonable for some applications. All subjects in the test felt quite comfortable using the system and quickly learned the user protocols (specified in the test instructions).

IV. DISCUSSION AND SUMMARY

This paper has described a speech recognizer used to control a repertory dialer system. The system uses speaker-dependent reference templates obtained from a training session prior to normal usage. Although we used two templates for each word in the vocabulary, it is practical and feasible to require just one by making the training algorithm a little more sophisticated. One way in which this could be done is to store the first replication of each word as a template and then compare the second replication to the first using the full recognition system. If the distance is below some suitable threshold, the first template is accepted and the system proceeds on to the next word. If not, the second template is stored temporarily and a request is made for a third replication of the word. This word is then compared with the first and second versions, and one of the templates from the best

match is chosen as the word template. Clearly, this process could be continued if no pair of words generate a close recognition distance.

The reliability and robustness of the system was demonstrated in a recognition test with six talkers and 4692 recognitions in which no recognition errors were made and only a small number of requests for repeats occurred.

The results presented here demonstrate that a task-oriented speech recognizer can be implemented in a reliable manner if one can take advantage of some of the natural constraints of the task, the vocabulary, and the recognizer.

REFERENCES

1. T. B. Martin, "Practical Applications of Voice Input to Machines," Proc. IEEE *64*, (April 1976), pp. 487–501.
2. A. E. Rosenberg and C. E. Schmidt, "Automatic Recognition of Spoken Spelled Names for Obtaining Directory Listings," B.S.T.J., *58*, No. 8 (October 1979), pp. 1979–1823.
3. B. Aldefeld, S. E. Levinson, and T. G. Szymanski, "A Minimum-Distance Search Technique and its Application to Automatic Directory Assistance," B.S.T.J., *59*, No. 88 (October 1980).
4. S. E. Levinson, A. E. Rosenberg, and J. L. Flanagan, "Evaluation of a Word Recognition System Using Syntax Analysis," Proc IEEE ICASSP-77, Hartford, Ct, 1977.
5. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-23*, No. 1 (February 1975), pp. 67–72.
6. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, (September 1973), pp. 1105–1118.
7. L. R. Rabiner and R. W. Schafer, "Digital Techniques for Computer Voice Response: Implementations and Applications," Proc. IEEE, *64*, No. 4 (April 1976), pp. 416–433.
8. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., *50*, (August 1971), pp. 637–655.