Application of Isolated Word Recognition to a Voice Controlled Repertory Dialer System

L. R. Rabiner
J. G. Wilpon
A. E. Rosenberg

Acoustics Research Department
Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

In this paper we describe a speaker trained, voice controlled, repertory dialer system. The main elements of the system include:

1. A real-time speech analyzer that detects the presence of speech on the input line, and analyzes the speech to give features appropriate for a word recognizer.
2. An isolated word recognizer that decides which of a set of words was spoken.
3. A voice response system to provide spoken commands to the user to guide the use of the repertory dialer system.
4. A dialer (simulated) to outpulse the desired telephone number.

The repertory dialer system is implemented on a minicomputer with a high speed array processor performing the real-time operations. The vocabulary for the system consists of 7 command words, 10 digits, and any number of names up to some specified maximum. Recognition is performed on one or more subsets of the vocabulary, depending on the state of the system.

To train the system the user is requested to speak each of the vocabulary words twice to provide reference templates for the system. Following 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 an isolated string of digits.

The system was tested extensively by 6 talkers (3 male, 3 female - 3 of whom were naive and 3 experienced users) over a three week period. A total of 4620 words were spoken and during the course of the test there were *no* recognition errors. A request for a repeat of a spoken word occurred about 2% of the time. These tests demonstrate the reliability and robustness of this voice repertory dialer system.

1. Introduction

Progress in isolated word 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):

1. Speaker trained or speaker independent.
2. Size and complexity of vocabulary.
3. Operating environment.
4. User training required to use the system.

The most accurate and reliable recognizers are those which are speaker trained, with small to moderate size vocabularies (10-50 words) of low complexity (i.e. all words are distinctly different), with a high quality recording system, a low background noise environment, and with a modest amount of user training. Such recognizers can reliably maintain accuracies of 99% or better across a wide range of users [1]. As some of the factors are degraded (e.g. larger vocabularies, more complexity, telephone inputs etc.) the reliability and accuracy of the recognizer tends to become worse. In order 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 persons name (letter-by-letter), the recognition accuracy on the letters is only about 70-80%; however for an 18,000 name directory, the name accuracy is close to 98% [3].

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

1. 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].
2. 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 that is consistent with the semantic constraints of a time of the day.
3. The use of a rejection threshold 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 2 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 of the above techniques to provide a very accurate and reliable recognition system.

In Section 2 we describe the basic operation of the dialer and in Section 3 we describe a series of tests which measured the performance of the dialer in a realistic operating environment. In Section 4 we provide a brief discussion of the results and summarize the main contributions of the work.

2. Operation of the Repertory Dialer System

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

1. A real-time speech analyzer that detects the presence of speech on the input line, and analyzes the speech to give features

CH1559-4/80/0000-0182$00.75

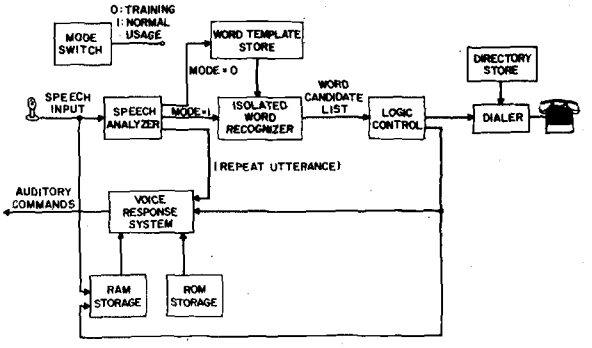

Fig. 1. Block diagram of the voice controlled repertory dialer system.

(frames of LPC (linear predictive coding) parameters) appropriate for the word recognizer.

2. 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.

3. 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 phases), and random access memory (RAM) storage (for the names in the directory).

4. A word template store for storage of the reference patterns for each word of the vocabulary.

5. A directory store containing the current set of repertory names and the associated telephone numbers.

6. A dialer to outpulse the desired telephone number.

7. Logic control to do the following:

   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).

8. 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 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 msec frames with a 30 msec 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 etc.), the user is requested to repeat the word. Complete training of the system consists of 2 replications of all the vocabulary words.

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	

Table I
Words Used in the Repertory Dialer Vocabulary

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 offhook preparatory to dialing a number. System responds with a single beep to prompt the user to speak a name in the repertory, or a string of 4 isolated digits.
HANGUP: Terminate use of the system; hangup 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 - either on the part of 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 offline (i.e. the telephone is onhook) 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

There are several key points that 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 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).

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"

Table II
Phrases Used by the Voice Response System

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:

1. SET 1 - 7 command words
2. SET 2 - 20 names, 10 digits, word STOP
3. SET 3 - 10 digits
4. 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 3 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.

3. Testing the Repertory Dialer

The voice controlled repertory dialer of Figure 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.

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 2 or 3 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 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 1 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 Figure 2 which shows plots of the

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) MATHEWS
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

Table III
Summary of Commands Used to Test the Repertory Dialer System

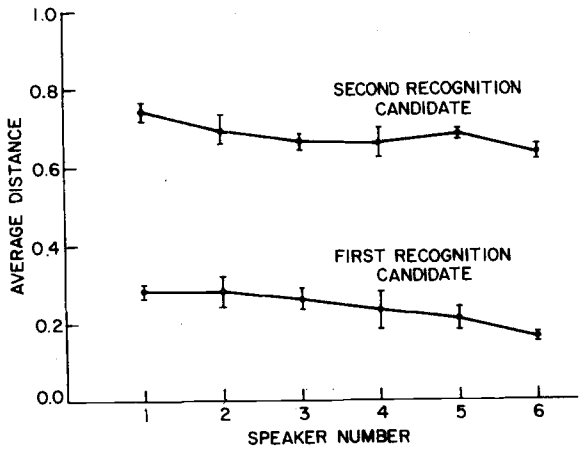

Fig. 2. Average distance as a function of speaker number for the first and second recognition candidates.

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%.

The only other statistic that was monitored during the tests was the average time for each test. On average, a complete test took about 12 minutes, or about 8-9 seconds per recognition, prompting, response and verification. Considering the amount of communication which 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 (they were specified in the test instructions).

4. Discussion and Summary

In this paper we have described a speech recognizer which is used to control a repertory dialer system. The system uses speaker dependent reference templates obtained from a training session prior to normal usage.

The reliability and robustness of the system was demonstrated in a recognition test with 6 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 advantages 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, Vol. 64, pp. 487-501, Apr. 1976.

[2] A. E. Rosenberg and C. E. Schmidt, "Automatic Recognition of Spoken Spelled Names for Obtaining Directory Listings," Bell System Tech. J., Vol. 58, No. 8, pp. 1979-1823, Oct. 1979.

[3] B. Aldefeld, S. E. Levinson, and T. G. Szymanski, "A Minimum-Distance Search Technique and its Application to Automatic Directory Assistance," submitted for publication.

[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. on Acoustics, Speech, and Signal Proc., Vol. ASSP-23, No. 1, pp. 67-72, Feb. 1975.

[6] P. Cumminskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," Bell System Tech. J., Vol. 52, pp. 1105-1118, Sept. 1973.

[7] L. R. Rabiner and R. W. Schafer, "Digital Techniques for Computer Voice Response: Implementations and Applications," Proc. IEEE, Vol. 64, No. 4, pp. 416-433, Apr. 1976.

[8] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., Vol. 50, pp. 637-655, Aug. 1971.