APPLICATION OF SPEECH RECOGNITION TECHNOLOGY FOR TELECOMMUNICATION SERVICES

Wilson Lo and A. Samouelian

Speech Technology
Technology Development Group
International Business Unit
OTC Australia

ABSTRACT - This paper presents the recognition results of two commercial PC based, isolated word, Speaker Independent Voice Recognition (SIVR) systems over the Public Switched Telephone Network (PSTN) with a vocabulary of 0-9 and several control words. A brief description of a pilot service called "World Time Information Service" which was developed using one of the SIVR evaluated is also described.

INTRODUCTION

The push button telephone provided the first Interactive Voice Response Service (IVRS) via tone dialing, in which users were prompted to select or enter information via the telephone push buttons. As the technology of the telephone advanced and the telephone exchanges became more sophisticated, the telephone companies started progressively to change the telephone sets from rotary to push button dialing, and as the exchanges were upgraded to accept tone dialing, whole telephone sets that were connected to the upgraded exchange area acquired the new feature of tone dialing. As the tone dialing became more widespread, primarily in the United States, so did dial up information services.

Over the last few years, developments in the speech recognition technology have reached a stage where implementation of voice interactive telecommunication services became possible. The technology has the potential of making IVRS more user friendly by allowing users to enter commands or select information from menus using voice commands. This is a more natural form of communication for humans. The technology also allows access to IVRS by telephone sets which are not connected to tone dialing, or in countries where the penetration of tone dialing is not widespread. The success of these services rely on good, speaker independent, speech recognition system over the PSTN and intelligent dialogue management.

To evaluate the current, commercially available SIVR systems, the Speech Technology group of OTC Australia evaluated two commercially available SIVR systems with a vocabulary of 0-9 and several control words. One recognizer was trained by the manufacturer on British English for a vocabulary set of 14 words (voc 1), the other was trained in-house on 100 speakers (70 male, 30 female) on Australian accented English on the same vocabulary set (voc 1) plus a further 5 words (voc 2). Both systems were evaluated over the PSTN.

To evaluate the dialogue management and the response of users to the IVRS, a pilot service called "World Time Information Service" was developed using one of the recognition systems that can be trained in-house.

PERFORMANCE EVALUATION TECHNIQUE

Ideally, to compare the performance of these different recognition systems identical speech files should be played to both machines. In this case, we can be sure of a fair comparison of the two systems. Unfortunately, this requires that the file format of the stored speech be suitable for both systems and as both the systems that were evaluated have automatic gain control and end-point detection algorithms based on signal level, sufficient background silence periods prior to and after the spoken word was required. For these reasons along with the requirement to develop special hardware and software tools to integrate the two SIVR systems, this method was not chosen.

Instead, it was decided to do "live" test, where callers were asked to make two identical phone calls to both systems saying a list of words in a specific order. The callers were unaware of which recognition board was connected to the specified telephone number. The recognition results of each caller were then logged and later evaluated. The callers selected were of half male and half female from various origins. All callers were OTC employees and were not necessarily computer literate nor familiar with speech recognition system dialogues. A description of the evaluation procedure follows:

The caller dialed the specified number and the respective system responded by answering the line (hook detect) and playing a pre-recorded greeting message to the caller explaining the recording instructions and then playing the prompting tone ready for the speaker to begin saying the words from the supplied list one by one. The program anticipated the words spoken by the caller in a specific order and if the current spoken word did not match the word expected then the caller was requested to say the word again. Regardless of whether the caller was successful on this second attempt the computer would request the caller to say the next word on the information sheet. The systems detected callers who spoke too soon, too soft or paused for too long and prompted them to repeat the word and callers were only permitted three such consecutive errors. After three such errors the system would assume a poor telephone line connection and the caller would be requested to call again later before being logged out.

PROBLEMS WITH "LIVE" TESTS

There were several problems with the "live" test, namely:

1. Not all callers made both calls to the machines from the same telephone, thus differences in line characteristics and background noise would effect the recognition performance.
2. Not all callers called both systems, instead some callers called one system only and asked another speaker to call the other system.
3. Because of the above two factors random errors inevitable from speakers were not duplicated onto both systems.
4. Some callers became out of sync with the expected word list and thus made consecutive errors. These callers were eliminated from the analysis.

ANALYSIS OF RESULTS

During the period of evaluation, both systems received over 250 calls. Any callers who made over four consecutive errors was marked as a bad call and the results of those calls were discarded. The overall performance results for both systems are shown in Table 1, while recognition results for each word in the vocabulary set is shown in Table 2.

Although the accuracy for each system is lower than claimed by the manufacturers, it is still substantially higher than was reported by R. Seidl (1990), when he evaluated Recognizer 1 that was trained by the manufacturer on American English. The performance of the British English vocabulary set is better than the American for Australian accented English. With careful dialogue management and the provision of confirmation feedback to the user, these systems could be considered suitable candidates for use in IVRS To demonstrate the capabilities of the speech recognition systems over

SIVR	*Number of callers*	*Vocabulary size*	*Number of tokens*	*Unrecognized words*		*Accuracy*
Pre-trained Recognizer 1 (voc1)	312	14	4,368	318	7.3%	92.7%
Recognizer 2 trained in-house (voc1)	255	14	3,570	346	9.7%	90.3%
Recognizer 2 trained in-house (voc2)	255	19	4,845	528	10.9%	89.1%

Table 1: Overall performance results of the recognizers

Word (voc 1)	*SIVR System*			
	Recognizer 1		*recognizer 2*	
	Number of Errors	Accuracy%	Number of Errors	Accuracy%
zero	26	91.7%	29	88.63%
one	0	100%	34	86.67%
two	21	93.27%	20	92.2%
three	19	93.9%	16	93.7%
four	22	93.0%	23	91.0%
five	25	92.0%	13	94.9%
six	12	96.2%	15	94.1%
seven	23	92.6%	25	90.2%
eight	24	92.3%	24	90.6%
nine	23	92.6%	26	89.8%
oh	43	86.2%	27	89.4%
cancel	31	90.1%	16	93.7%
help	31	90.1%	44	82.8%
stop	8	97.4%	34	86.7%
Additional Words in voc 2				
naught	-	-%	32	87.5%
yes	-	- %	21	91.8%
no	-	- %	95	62.8%
next	-	- %	11	95.7%
start	-	- %	23	91.0%

Table 2: Recognition results for each word in the vocabulary set

the PSTN, a pilot service called the "World Time Information Demonstration" was developed using the in-house trainable speech recognizer.

WORLD TIME INFORMATION SERVICE

The World Time Information Service demonstration allowed the user to nominate a destination country, and the system responded with the time and day in that country at the time of call, the call charge rate per minute and the country code information. This service simulates current operator assisted service of directory inquiries.

Recognizer 2 was trained on 30 speakers which consisted of mainly English speaking Australians with the majority of the callers being male. The vocabulary set of 20 words consisted of 16 country names and their variations, e.g. U.S., United States, America, U.K. and England. The speech output or system response was provided by the concatenation of message (vocabulary) elements using digitized speech.

SYSTEM CONFIGURATION

The configuration of the hardware is shown in Figure 1.

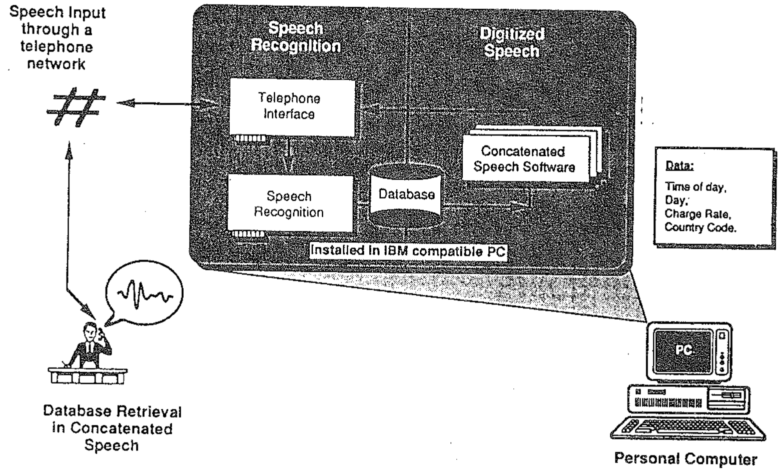

Figure 1: Hardware configuration for the World Time Information Service

When a call was received, the user was greeted by a greeting message and a request to nominate a country name. If the country name was recognized by the system, it prompted the user for confirmation. Otherwise the utterance was treated as an invalid option and the request procedure was repeated. If the system failed to recognize the nomination after three attempts, the call was transferred to another line for connection to an operator assisted position (Help Desk).

After the destination has been confirmed, the system would scan the database and extract the appropriate information on the selected country and respond to the user in the form of concatenated digitized speech.

PERFORMANCE OF THE DEMONSTRATION

The system has drawn positive response in its initial trials. It is currently planned to set up a field trial system for 40 most used overseas call destinations by international callers.

The collection of new speech database of these country names from the general population has recently been completed. The database consists of over 500 speakers (50% male, 50% female) from various mixtures of 1st and 2nd generation Australians to new migrants from European and Asian countries. The database will be divided into training and test sets and recognizer 2 will be used for this service.

CONCLUSION

The evaluation results of the two commercially available SIVR systems indicate that although the recognition accuracy is not in the high 90's as claimed by the manufacturers, through careful dialogue and user interface, it is possible to consider these systems for the development of IVRS for telecommunication services using speech recognition technology.

ACKNOWLEDGMENTS

The permission of the Managing Director of OTC Australia to publish this work is gratefully acknowledged. The views and opinions expressed in this work are those of the authors and do not necessarily imply OTC policy of future service offerings.

REFERENCE

Seidl R. (1990). "The Application of Speech I/O technology to Interactive Telecommunication Services", Third Australian Int. Conf. on Speech, Science and Technology, November 1990, Melbourne, Australia, pp 480-485.

THE ROLE OF HUMAN FACTORS TESTING IN SPEECH TECHNOLOGY

Dr Elizabeth Bednall
Josephine Chessari
Human Factors Team
Telecom Research Laboratories

Three studies are reported which focus on human factors issues in three speech-based telecommunications products. The purpose of the paper is to describe how human factors methodology can be applied in different ways. The first product shall be referred to as System A, the second as System B and the third as System C. The methods used provided valuable information regarding the human-computer interface for each system. These included:

1. user needs analysis
2. heuristic evaluation
3. observation of users interacting with the system while completing typical tasks
4. measurement of performance of the system
5. interpretation of questionnaire data
6. testing of product managers on the tasks

The results of these studies clearly illustrate the need for human factors testing to be incorporated into the design process. Such testing ensures a cost-effective way of optimizing usability and customer satisfaction when a product is finally released into the market place.

1. INTRODUCTION

It would be easy to assume that because speech recognition systems offer a "natural" interface between humans and computer systems, the design of such interfaces is just "common sense" and requires little effort. Paradoxically, it may be more difficult to get such an interface right, given well learned conventions in human speech behaviour which may or may not translate easily to the human-machine interface.

Human factors specialists ideally have a background in cognitive psychology and are well placed to study the behaviour of humans whenever they are required to interact with computer-based systems. The objective of any human factors study is to discover problems the human operator is likely to experience when using a given system. Solutions can then be recommended which take into account the cognitive capacity of the typical user.

There are a number of tools which human factors specialists use to shed light on problems with a particular interface. Usually, a combination of such tools provides the best view of design problems from the human factors perspective. Human factors studies may be carried out quite early in the design process, ideally in a test, modify, re-test (iterative) fashion, to ensure that the best possible system from the user's point of view is created.

It will become clear that an important component in deciding how to test a particular product is to establish, in advance, who the intended market is likely to be. In particular, despite the fact that the Systems A and B are nearly identical in many aspects, the target market is completely different. It is essential, then, to place the product in its correct context, before trying to establish how well it works.

2. "SYSTEM A" STUDY

System A is a telecommunications product which requires the user to dial into the service using a telephone. The user then enters a total of 20 digits to connect to a set of recorded voice announcements which instruct the user on how to proceed. Digit entry can be achieved in two ways. On all phones, the user can enter the number by saying the digits into the phone. The