ENGINEERING A SPEECH-CONTROLLED VOICE-MAIL DEMONSTRATION SYSTEM OPERATING ON THE TELEPHONE NETWORK

A.J. Hunt, P.C.B. Henderson, A. Samouelian*, J.M. Song, and R.W. King

Speech Technology Research Group
Department of Electrical Engineering
The University of Sydney

* Speech Technology
Technology Development Group
International Business Unit, OTC Australia

ABSTRACT - Engineering real-time speech recognition into services to operate over the telephone network requires more than good speech recognition. This paper addresses the design and integration issues of such a system; a speech controlled voice-mail demonstration system. This system integrates a PC-based telephone interface card, DSP signal processing, Sun-based isolated-word speaker-independent HMM recognition, an X-window user interface, prompt generation and system control. The system provides a base for evaluating the effectiveness of speech-controlled applications as well as being a pilot for investigating the difficulties in implementing real-world speech technology systems. This system also demonstrated the robust performance of a recently developed HMM package.

INTRODUCTION

In recent years speech recognition systems have reached a sufficient sophistication and robustness to allow them to be used in real-world applications. However, the commissioning of such applications requires taking into account several issues. Human-factors issues affect user acceptance and system effectiveness. Widely varying conditions encountered outside the laboratory environment affect performance, and in particular, applying speech recognition over the telephone network needs to take special account of the network's imperfect nature. So, engineering speech recognition into real-world services and systems requires more than good speech recognition.

This paper addresses the design and integration issues of such a system, a speech-controlled telephone-based voice-mail demonstration. The system integrates a PC-based telephone interface, DSP-card based signal processing, Sun-based isolated-word speaker-independent HMM recognition, an X-window graphical user interface, prompt generation and system control. The HMM technology utilised in the system is described in detail elsewhere in these proceedings (Song & Samouelian, 92).

This paper also discusses experience gained in performing the implementation of the system, and ideas for enhancing the system are introduced. The system demonstrates the robust performance of the HMM recognition and the utility of a good speech interface.

USER INTERFACE

The primary task was to put together an operational demonstration using HMM recognition software which had been developed by Song & Samouelian (1992). The demonstration would meet the complementary demands of evaluating the HMM technology, and of investigating the issues arising from applying such generic speech recognition technology to a real-world application. The demonstration needed to be as real as possible to see whether a speech-based system would be acceptable to users.

A voice-mail system was chosen because it was a realistic application which leant itself well to a speech-control interface with a small word set. Two restrictions were made to ensure the task was possible within a short development period. For demonstration purposes, the system used pre-recorded messages, and no security access mechanism was implemented.

The voice-mail interface thus simulated access to incoming messages and provided the ability to save messages and play them back at a later time The control mechanism had to provide the ability for a user to select messages to be played. ability to control the playing of messages (restarting, stopping, continuing, and so on), and the ability to save and delete messages. The system also included context-sensitive help at any point in the demonstration.

The interface used 19 control words - YES, NO, STOP, START, CANCEL, NEXT, 1, 2, 3, 4, 5, 6, 7, 8, 9, OH, ZERO, NOUGHT and HELP. The words OH, ZERO and NOUGHT were not required for the current implementation of the system but were included to test the recognition of highly confusable word-pairs (e.g. OH - NO) in real-world conditions.

The user interface prompts were kept simple to provide a fast interface. No checking prompts were provided (an example may have been "Are you sure?"). This did not significantly detract from the robustness of the system, but did make the interface more "user-friendly". The prompts were designed through multiple iterations with user feedback to make the interface as natural and intuitive as possible. This style of development was found particularly effective and was supported by appropriate software, described later in the paper.

A novel feature of the demonstration system was the handling of simultaneous speech input and output. It was possible to speak a voice command while a prompt or a recorded voice message was being played. The main advantage was that users could stop the playback of possibly long recorded messages. An added advantage is that familiar users could operate the system more quickly and efficiently by pre-empting the prompts. It is an improvement over most existing systems which require users to speak only after a beep. However, a difficulty is introduced due to cross-talk on telephone channels - this issue is discussed later in the paper.

By providing a simple intuitive command set, a simple dialogue structure and flexibility in the user input timing, we aimed to provide an interface to which users would respond comfortably.

SYSTEM DESIGN AND INTEGRATION

Figure 1 shows a block diagram of the voice mail demonstration system. The major hardware components are the PC with a telephone interface card, SUN SPARC II with a DSP32C card and running X-windows and an ethernet connection between the two.

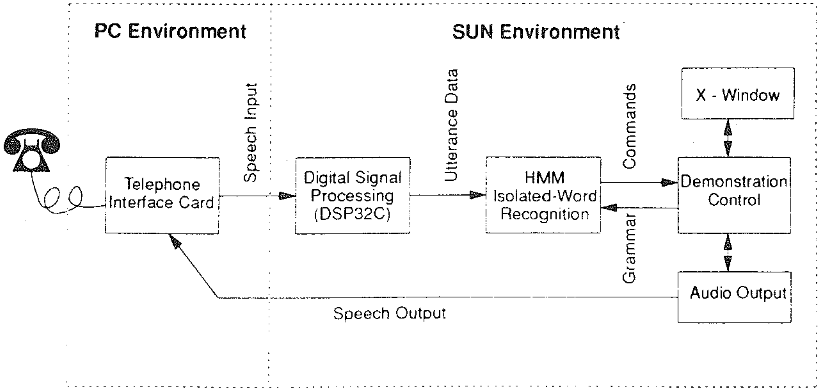

Figure 1. System Block Diagram.

The major software components are the telephone interface control on the PC, software to connect the PC and the Sun over ethernet, signal processing software on the Sun and DSP card, the HMM recognition software, the X-window display manager, the prompt generator and the control software. Each of these sub-systems is described briefly below.

Telephone Interface

The telephone interface card provided four telephone lines. One line was used in this demonstration system. The software on the PC accepts an incoming call and then enters its sampling mode. The data is μ-law encoded and sampled at 8kHz giving a data rate of 64kbs. No signal processing was performed on the PC or on the telephone interface card, instead a continuous stream of incoming data was sent over ethernet to the Sun for signal processing and recognition.

Signal Processing

The Sun accepted a continuous stream of incoming sample data from the PC. The data was then sent to the DSP card which performed μ-law decoding and produced front-end data frames every 10msec and in real-time.

Next, end-point detection was performed. As the system was isolated-word based a simple algorithm could be used to find the utterances in the incoming signal stream. A peak in the energy values for a minimum period was used to trigger the start of an utterance. A longer period prior to the trigger was included so that leading low-energy phonemes (e.g. fricatives) were not lost. The end-point was determined as being a fixed time after the last high energy frame, with end-point shifting back in time each time the energy peak re-triggered. This re-triggering meant that multi-syllabic words would keep re-triggering the system. The triggering level and the various timing periods were determined empirically. They were set so that unwanted noise (e.g. a bump caused by someone placing the telephone mouthpiece down on a table) would not cause a false trigger.

HMM Recognition

Once an utterance had been found the HMM process attempted to recognise the utterance. The system had a 19 word vocabulary and a silence model which identified and ignored leading and trailing silences around an utterance. The demonstration could be run in two modes, either with or without context restriction of the word choices. The list of legal words at any point in time was provided to the HMM recogniser by the demonstration controller.

The accuracy of the HMM recognition in various operating conditions is discussed in Song & Samouelian (1992). The time taken to perform HMM recognition processing was insignificant compared with other processes in the system.

X Display

The demonstration included an X-window based display manager which illustrated the internal operations of the demonstration. The display was useful as a feedback mechanism to users of the demonstration in the laboratory environment, to developers as an analysis tool, and to new users as a training aid. Four windows were maintained. The first window covered the status of the demonstration including such information as details of the current connected call, and the availability of recorded messages. The second window displayed full details of available messages and updated as these messages were played, saved or deleted. The third and fourth windows illustrated the operation of the HMM recognition. A list of legal words was displayed along with a display of the results of the recognition of the last utterance, including the relative probability of occurrence and the ranking of candidate words. These four windows were a very useful adjunct to the core demonstration capabilities and were of interest to both technical and non-technical users.

Audio Output

The system provided both audible prompts to the users and audible responses to user commands. As discussed earlier, these prompts were designed through multiple iterations. The prompts were pre-recorded in a high-quality environment. The simulations of the recorded voice messages were recorded using the low-quality audio capability of the Sun SPARC workstations which provided a realistic feel to the demonstration. Dates, times and several other prompts were produced by concatenation of speech segments. All prompts were output through the standard audio device provided on SUN SPARC workstations which was hard-wired onto the telephone interface card.

Control

Underlying the voice mail system was control software which determined the appropriate response to each command, controlled the X-window interface, set the vocabulary restriction for the HMM recogniser and controlled the audio output. The system was designed as a state transition system with incoming events (e.g. spoken commands) producing desired actions, and, if needed, a state change. A state-based system is a flexible base for introducing enhancements. Also, this flexibility proved particularly useful for the iterative process used in developing the user interface, as described earlier.

DISCUSSION

The system showed that the HMM recognition software that had been developed could be used in a real-world system. The recognition was robust under differing noise conditions and responded well to a very wide range of speakers with different sex, accent, and voice characteristics. The system showed also that a basic speaker-independent speech recognition system could be used as a natural interface to a real-world system. The interface was simple to use and naive users operated the system with almost no training. The general response to the demonstration was positive.

The system response time varied between 1 and 3 seconds. Users found the response time satisfactory. The primary delay in the system was the communication between the PC and Sun environments. This delay will be reduced by the changes in environment which are discussed below.

The development of the system did show up some inadequacies of the environment. First, the integration of the PC and Sun (UNIX) environments was troublesome - a problem which is not unique to this system. Reliable communication between the two environments was difficult to achieve. A second problem is that UNIX is not a good environment for the development of the real-time response and processing software. However, UNIX did provide powerful support for much of the development of the system. In future, systems would be better if implemented in one environment or the other.

The imperfect nature of the telephone system introduced a particular difficulty. The cross-talk between incoming and outgoing channels on the phone connection caused problems. Telephone lines carry a single pair of lines, and precise balancing was required to separate the incoming and outgoing signals. Poor line balancing on some connections lead to the output signal feeding back into the incoming signal. This caused some false triggers of the recognition, and added noise to some user utterances. The system performance was intermittently affected, but was still acceptable. The problem can be addressed in a few ways. The balancing can be adapted to each incoming line - this would require special hardware. Dynamic filtering can be used to cancel the feedback signal. The user can be required not to speak over output prompts, though this would work against the original intentions of the interface design. Work on this problem is continuing.

CONCLUSIONS

The demonstration system showed that good speech recognition can be transferred into a working real-world system if the design is well thought through. The HMM recognition modules must be trained on real-world data. The user interface must be simple and intuitive for users to respond well to the system, and this requires careful design. The system had satisfactory response time and recognition accuracy. However, additional work needs to be done on dealing with the imperfect communication environment provided by the telephone network.

ACKNOWLEDGMENTS

This work was carried out under contract to the Speech Technology, Technology Development Group, International Business Unit, OTC Australia, 255 Elizabeth Street, Sydney, NSW, 2001, Australia. 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.

REFERENCES

Song, J.M., & Samouelian, A. (1992) *A robust speaker-independent isolated-word recogniser over the telephone network based on a modified HMM approach*, SST-92, elsewhere is these proceedings.

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