HUMAN-COMPUTER SPEECH COMMUNICATION SYSTEMS

C Wheddon

Speech and Language Processing Division
British Telecom Research Laboratories

ABSTRACT

The principal means of human communication is speech, this modality is now replicated by computer systems that are able to hold a limited but useful conversation with the user. Systems in operation and under development are described.

INTRODUCTION

Speech dominates the world's telecommunication systems, it accounts for over 800 billion telephone calls a year made from a vast variety of telephones. There is a familiarity with the telephone terminal which pervades most social and economic groups and extends over most age groups. This telephone culture is now converging with the explosive growth of electronically stored information. Computer databases are increasingly being used to store information on timetables for buses, trains and airplanes, as well as financial, commercial and medical data.

Telephony access to these databases can be achieved by communicating with a telephone operator trained to use a computer terminal. This method of indirect access is limited by the number of operators available to answer calls and deal with inquiries, this imposes limits on the transaction times and availability. The ubiquity of the telephone system therefore presents an opportunity for direct access to computer databases by speech, providing the problems of human interaction with computer's can be overcome.

SPEECH TECHNOLOGIES

The recognition and generation of speech by computers involves many complex processes that are achieved routinely and without effort by humans. Current recognition technology allows a few hundred words to be understood if each word is spoken in isolation or a limited number of fluent phrases if articulated in a deliberate manner. However, despite the vocabulary limitation of current computer recognition systems, useful applications can be achieved especially if they can be designed to operate over the telephone network.

Speech output by computers can be achieved either by converting textual information through a text-to-speech system which uses a set of linguistic rules (grapheme to phoneme) to produce parameters which drive an electronic vocal tract model to produce speech. The text-to-speech system is able to convert un-restricted text into speech, although the quality is such that most users require a period of exposure before they become tuned to the speech.

Alternatively speech can be coded directly from the waveform of natural speech or synthesised from parameters extracted from the speech waveform. In either case redundancies are removed from the speech signal to achieve

low data storage; quality is preserved as close to the original speech as possible. The main disadvantage with speech output derived by these methods is that the vocabulary must be known in advance, which limits the flexibility of the application.

TELEPHONE RETRIEVAL OF DATABASE INFORMATION - CAESAR

Before the recent advent of speech recognition over the telephone network, interactive database retrieval systems relied on the message repertoire of the telephone keypad. These systems provided voice guidance to the user who then interacted with the system by keystrokes to obtain the required information from the database.

A system based on the user interacting with a large database via the telephone keypad has been produced and is now in operation as part of the new British Telecom Customer Service System (CSS). The CSS is an integrated database with national coverage via 29 linked district computer centres. A typical district installation might be:-

1M customers	50Gbyte disk storage
2K terminals	350K transactions per day
140M database records	

The system developed for linking into the CSS database is called CAESAR. It provides an intelligent interface between the Public Telephone switched Network (PSTN) and the computer database. CAESAR has been designed as a programmable voice response unit capable of reformatting information from computer screens into speech. One of the first applications of the CAESAR system enabled installation engineers to access to network records at any time day or night.

On accessing the system and gaining entry to the database the user can specify the information required by means of keying in single character responses in reply to CAESAR's spoken guidance. The information is then read from the database and converted in speech by a text-to-speech synthesiser.

It was extensively trialled before it became operational with users able to influence the design of the entry procedures and the type and format of the spoken message responses. Interestingly initial triallists did not report adverse reactions to the text-to-speech system.

The general arrangement of the Caesar architecture makes it easy to adapt to other applications and is shown in figure 1.

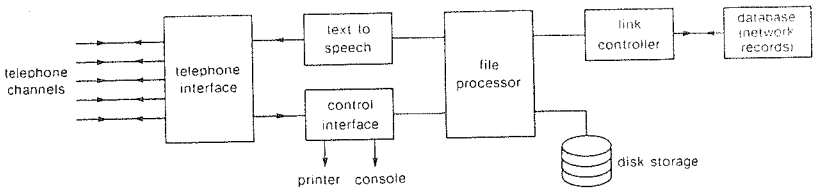

Figure 1. Caesar Base Unit

VOICE OPERATED DATABASE INQUIRY SYSTEM - VODIS

The telephone keypad can only cater for a limited message repertoire and is therefore limited in its scope for human-computer interaction. VODIS is a voice-operated demonstration system that is intended to enable a novice user's access to a database - such as train-timetables.

The main components of the voice operated database inquiry system are shown in figure 2.

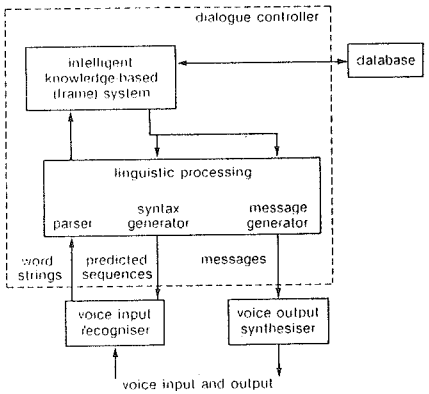

Figure 2. Voice Operated Database Inquiry System

In operation the system starts by asking the user a question and predicting the likely responses. The intelligent knowledge base, which runs the linguistic processor on a mini-computer, uses a frame-based language to build up knowledge on the inquiry relative to the particular goal of providing train timetable information.

The procedures are data driven ie. they are evoked as data becomes needed or is input. As a consequence the path of the dialogue is only loosely controlled and is determined by the users responses. Depending on the status of the dialogue, the knowledge base requests the message generator to construct an appropriate question and at the same time primes the syntax generator on the expected range of meaningful replies. An adaptive strategy is used such that when the recognition (of the fluent spoken passage) is good, the inputs are relatively unconstrained, but when the performance is poor the questions are made more specific.

To test potential user's reactions to the voice operated inquiry system, a total of 32 members of the general public participated in a trial. None of the subjects had ever used the system before and they were not given any specific instruction about using VODIS apart from the procedure to initiate their conversation with the system and to give the general

scenarios for the task of finding departures, arrivals and journey times for a number of destinations on the east coast rail link in the UK. The results showed that for male users the completion rates varied from 60% to 100%, but were poorer for female users.

Further development of the VODIS system should enable relatively inexperienced users of technology to simply speak their requirements for: trains, theatre bookings, telemarketing services and for a vast range of database inquiries in the next few years.

AUTOMATED TELEPHONE BANKING - ATB

BT has developed a telephone banking system for public trial with the Royal Bank of Scotland using voice recognition and speaker verification techniques that allows customers to make enquiries via the ordinary telephone. By speaking to their bank's computer, authenticated callers can obtain the balances of their accounts, details of their last few transactions, request a cheque-book or statement, pay pre-designated bills and transfer monies between accounts.

On phoning the system the caller is required to speak an identification number followed by a password (neither of which need to be kept secret) and a system-generated word.

The password and the random word are checked by the speaker verification unit against features previously stored by the customer, thereby providing a high degree of security.

The current system has both public and derived telephone service connection from 4 sites within the UK. It employs telephone switching, local area network (LAN) technology, advanced voice systems technology, audio recording of transaction and an intelligent help desk. Nearly 400 people are registered users of the system and by October, when the trial finishes, results of the users' responses will be known and assessed.

SPEECH AUTOMATIC LANGUAGE TRANSLATION

Language translation by humans is hard. It is even harder when machines attempt to do this by detecting, recognising, understanding and translating spoken phrases. A system under development by BT attempts to overcome some of the difficulties of translation by firstly limiting the range of discourse, then by keyword spotting of selected linguistic features and then finally by detecting the separation of phrases. An accurate translation of spoken phrases is achieved by using a structured phrase book approach.

Although the system is phrase-book orientated it is extremely robust to errors. The system does not require the user to know the exact contents of the phrase-book - just that the spoken phrase is close enough to one of the stored phrases. For example:

PLEASE WOULD you reserve me A single ROOM WITH bath

NB. The words in capitals represents the words spotted by the recognition system.

The closest phrase in the phrase-book:-

I WOULD like TO book A single room WITH bath PLEASE

which translates into:-

FRENCH Je voudrais reserver une chambre pour une personne avec salle de bains.

GERMAN Ich mochte ein einzelzimmer mit bad reservieren.

SPANISH Desearia reserver una habitacion individual con bino.

Figure 3 illustrates the general arrangement of the hardware.

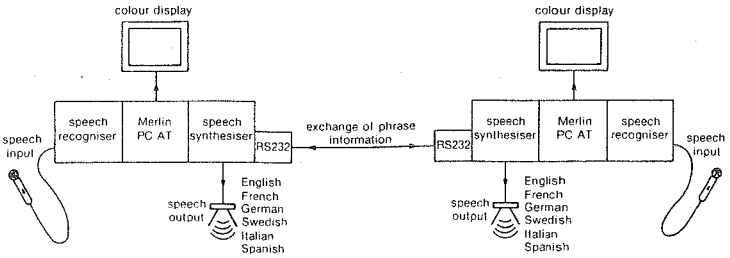

Figure 3. Simplified Block Diagram of Prototype Hardware

FUTURE RESEARCH DIRECTIONS - MAN-MACHINE COMMUNICATION SYSTEMS

Speech, vision and language processing has progressed to a stage beyond which significant advances can only be made by the use of neural models. These neural models or networks use a particular topology and learning rules for the interactions and interrelations that are based on the connections of the human brain. Neural networks process information in a number of different ways and are able to exhibit useful properties such as association, generalisation, differentiation and optimisation which are forms of self-organisation that cannot be easily achieved by conventional digital computers. BT, in common with other advanced research organisations, are currently working in this new and exciting R&D domain. Prototype speech recognition and image recognition devices based on techniques of the multi-layer perceptron have been developed as research models.

The future of man-machine communication systems based on neuronal networks hold the prospect of human like performance in a number of information processing tasks that will initially be used as better human-computer interfaces.

REFERENCES

McCulloch, N. Ainsworth, W. A. & Linggard, R. "Machine Translation of Speech", British Telecom Technology Journal, Vol.6 No.2.

Steer, M.G. & Stentiford F.W.M. "Machine Translation of Speech", British Telecom Technology Journal, Vol.6, No.2.

Wheddon, C. "Speech Communication", British telecom Technology Journal, Vol.6, No.2.