REVIEW OF BRITISH TELECOM SPEECH RESEARCH: OBJECTIVES AND STRATEGY, 1986.

R. Linggard R18.3.3
British Telecom Research Laboratories

ABSTRACT - Speech Research and Development at British Telecom Research Labs is carried out in Division R18. This unit, of about 90 Scientists, Engineers and Technicians, undertakes long-term research, product development, and design of prototype equipment. This paper describes and discusses the objectives and strategies of British Telecom Speech Research in 1986, and outlines some of the projects now in hand.

INTRODUCTION

R18, the Speech Recognition, Synthesis and Coding Division, usually known as the Speech Technology Division, is the unit in British Telecom Research Labs charged with specifying and designing new speech products for BT's Operating Divisions. In addition, the Speech Technology Division also builds prototype equipment for evaluation purposes. Actual production of speech products and processing equipment is carried out by BT's subsidiaries, or licenced out to third parties.

Research and Development of specific equipment is sponsored by the BT Operating Divisions, who play a large part in identifying new products. They also supply market research information for potential products, and conduct appraisals of prototype equipment. Background, (strategic) research is mainly directed and funded centrally, though some longer term projects are inspired and financed by operational requirements.

The short-, medium- and long-term research projects are, to a large extent, interrelated and interdependent. The shorter term (0 - 1 year) research is directly motivated by current development projects, and unresolved problems from this may inspire medium or longer term research. However, medium-term (1 - 3 years) research is mainly driven by the specific requirements of new products, and the long-term (3 - 10 years) research is aimed at fulfilling the requirements of future speech products. Moreover, since these future products are themselves somewhat hypothetical, their actual definition may also be the subject of research projects.

EXISTING PRODUCTS AND EQUIPMENTS

The future, all-digital, telephone network in the UK will be based on 64 kb/s channels. Digital Telephone Exchanges are already in service and this has motivated research into encoding/storing/processing/decoding of speech signals in digital format. Speech-based equipment designed for this system is now in use, providing the Speech Technology Division with an invaluable background of field experience.

One of the first digital speech systems was the Automatic Announcement Service (AAS) designed for the digital telephone exchange of BT's System X. The announcements for AAS are based on words and phrases of natural speech, digitally encoded and stored in a large memory. These are retrieved and joined together under computer control to form the required announcements. By careful choice of vocabulary, and by storing several versions of some items, each with a different stress pattern, the resultant speech can be made to sound very natural. The disadvantages of AAS are in the time taken to prepare the vocabulary and the difficulty of adding new items. Since the speech was originally provided by a human female speaker, her availability and the long-term stability of her voice, presents problems in up-dating and developing the service. It was this and similar experience which led to our present research programme on speech synthesis from text.

Encoder/decoders (codecs) to transmit speech at rates less than 64 kb/s are useful in multiplexing single digital channels. R18 has designed a range of such vocoders, going from 32 kb/s down to 16 kb/s. An equipment based on these is the QUADMUX which multiplexes four speech channels onto a single 64 kb/s line. Another. DATAMUX, permits data channels at 16 kb/s to be multiplexed with speech data and transmitted on a 64 kb/s line. A more recently developed 8 kHz codec, is being tested for use as a secure phone. An alternative to reducing the speech data rate, is to increase its bandwidth, and a system to transmit 7 kHz speech over the standard 64 kb/s line has been designed and constructed.

Another existing speech processing machine is the TALON equipment, which permits digitised speech to be stored and accessed, for recorded message applications. The messages are digitised and stored on a large disc with buffer memory for each of 32 channels. The total amount of stored speech is 4 hours, which can be apportioned to the channels on a flexible basis. The play-back of messages can be broadcast or allocated to individual channels. A significant advantage over earlier, mechanical equipment is that the played-back message can be set to start at the beginning. This is a very important feature for many new types of network service.

Information-providing systems based on the TALON equipment, require a separate number to be dialled for each separate message. A more flexible system would permit a single number to access the whole range of information available, with the user then selecting individual data items via some form of interaction. Technically, the simplest form of communication between the user and a computer information service is via the MF4 (multi-frequency) key pad. Unfortunately, most customers do not yet have telephones incorporating this device. However, an in-house information service (CAESAR) for BT line-engineers has been built, which gives out line-connection data from a computer data-base. The line-engineer accesses the data-base via an MF4 key pad, using a sequence of numeric codes. The system outputs the required information via a text-to-speech synthesiser. Since the output text is rich in proper names, the

system is designed to revert to a word-spelling mode if so requested. Even in this simplest of forms, the problems of dialogue control become manifest, and have become an important field of research for us.

Obviously, the most natural and convenient interactive media for use over telephone lines, is speech itself. However, reliable, speaker-independent recognisers are not yet available, and speaker-dependent machines are severely limited in their applications on a public telephone network. One service which has been implemented, using speaker-dependent recognition, is a voice-controlled game. The game takes the form of an adventure story, in which the caller himself is the hero. Voice input enables the user to steer the adventure in one of several directions at each juncture in the story, in order to guide the adventure to a successful conclusion. In the preface to the game, the rules are explained and the user is asked to repeat the control words. These examples of the user's voice are then used to train the recogniser. Because the control words are used in citation form, the recognition rate is high, and the games have been very successful.

An application in which speaker-dependence is an advantage rather than a liability, is in speech-dialling. A speech-controlled telephone ASCOT, has been designed and constructed, which allows the user to access the netword by spoken commands. Initial tests indicate that the most useful mode for this device, is as a repertory-dialler. That is, the user speaks the name of the person to be called, rather than the number. The machine can store about 50 different numbers, each accessed by a unique name. Since numbers and names must be entered into the machine for each individual user, the training of the speaker-dependent recogniser can be carried out at the same time, and is not an extra chore. This telephone is not yet commercially available.

A more recent development in the Speech Technology Division, is a single-board, Text-to-Speech sub-system, which converts ASCII characters at 150 bits/s into synthetic speech. The synthesiser is a digital, parallel-formant type, and uses BT's own digital filter chip, controlled by a 68000 processor. There are 200-300 spelling rules, about 20 prosodic rules, and a 2000 word exceptions dictionary. Extra exceptions can be included to suit specific applications. In addition, this unit can be used to reproduce high quality, parameterised speech, which is supplied as a custom service.

NEW APPLICATIONS PROJECTS

Perhaps the most ambitious of the new applications projects in the Speech Technology Division, is VODIS (voice operated data base inquiry system). This is being carried out as part of the Alvey Initiative on Information Technology. A train time-table enquiry system has been chosen as an example of an information service which could be automated. The target is to permit a user to find

out train times between specified destinations using a fairly natural dialogue. Although the ultimate system will eventually require the recognition to be speaker-independent, a speaker-dependent recogniser is being used to facilitate research into the system design. The output from VODIS is speech synthesised from text. Experience on this project has fed through to both speech synthesis and speech recognition research. With a service of this sophistication, system organisation becomes an important part of the project, and one of the chief motivations of VODIS is to investigate the problems of dialogue control.

As well as investigating voice-based systems for use on the public telephone network, the Speech Technology Division is also researching advanced office information systems linked to the network. An important example of this research is the Advanced Telephone Answering Machine. This device is essentially a telephone minder that you switch on when leaving the office. An incoming call is greeted by a synthesised voice explaining that you are not in and asking the caller if he would like to leave a message. If the answer is "yes", then the message is recorded. If the answer is "no" then the machine initiates a dialogue to persuade the caller to at least leave a number. The advantage of this equipment over conventional answering machines is that a high proportion of callers actually do leave some kind of message. The recognition vocabulary for this machine is quite small, but it must be speaker-independent. Once again, dialogue control turns out to be a critical aspect of the design.

The facilities of the ASCOT repertory speech-dialler, are being built into potential new products. By multiplexing several basic ASCOT machines and using a disc memory, a voice-dialling service for private exchanges is being developed. Because of the multiplexing factor, a few recognisers can service a large number of lines, so that the cost per line is quite small. The extra cost of this service turns out to be only a fraction of the cost of the exchange.

A situation in which voice-dialling has considerable advantages, is in mobile telephones, and the Speech Technology Division has incorporated its speech-dialler electronics into a mobile phone. In an "in-car" situation, hands-free dialling is an important facility, especially as it now seems likely that there will be legislation to ban the use of hand-held phones whilst driving. By including AGC and noise cancellation, the basic speaker-dependent recognition algorithm has been adapted to cope with the background noise. This new system, known as TOPAZ, was developed in collaboration with British Telecom's Mobile Phone Division, and is about to go into production.

FUTURE PRODUCTS AND STRATEGIC RESEARCH

Rather than itemising the myriad potential future products based on speech technology, it is more instructive to identify generic product types, and to link these to a schedule of strategic research areas. The diagram in Fig. 1 attempts to do this.

ACKNOWLEDGEMENTS

The help, advice and encouragement from colleagues in the Speech Technology Division is gratefully acknowledged. In particular I owe thanks to; I. Bruce, P. Challenor, N. Condick, D. Gibson, P. Hughes, D. Johnston, P. Millar, C. Southcott, F. Stentiford, J. Tuppen, and C. Wheddon.

The author also wishes to thank the Director of Research, D. Merlo, for permission to publish this paper.

FIGURE 1.

FUTURE PRODUCT TYPES AND STRATEGIC RESEARCH AREAS