Review of speech technologies for telecommunications*

by F. A. Westall

Speech is the easiest and most expressive means that people have for communicating with each other. This paper reviews the key speech technologies and discusses the associated telecommunications applications and technical challenges. It concludes with some personal predictions about future trends and opportunities in this important, exciting and far-reaching field.

1 Introduction

As people have been communicating with each other with some form of speech for at least 50 000 years, it is not surprising that most of us take it for granted. Yet speech is a highly complex process and even the simplest sentence contains a world of information besides its literal content: it conveys the essence of human emotion, mood and personality. Getting a machine to generate such subtle speech is somewhat of a challenge; making machines which can fully comprehend speech is even more daunting.

Despite these challenges, worldwide interest in speech technology is growing at an unprecedented rate. Telephone speech is core telecommunications business, currently accounting for over 90% of revenues for the public telecommunication operators. It is also the primary access medium to 26 million subscribers in the UK telephone network, and to potential revenue growth opportunities from a base of around 700 million telephone users worldwide.

Many technical, commercial and regulatory factors contribute to the growing interest in speech processing. But it was the advent, in the late 1970s, of the single-chip digital signal processing microcomputer which helped to convert research into practical, cost-effective systems. This has resulted in the current upsurge of interest in new telecommunications applications for speech technology. Since the advent of the transistor, device complexity has been doubling on average every two years or so. It is interesting to note that if aeronautical engineering had progressed at the same pace as device technology, then it would now take under a minute to fly across the Atlantic!

Speech processing applications are today in the vanguard of this revolution, just as modems were a few years ago. The latest generation of digital mobile phones, aeronautical phones and multimedia terminals depends on speech coding. Speech recognition is used commercially in interactive voice systems, such as network-based voice messaging, telephone banking, and train timetable and directory enquiries. Speech echo cancellers are regularly used on international calls to facilitate two-way simultaneous speech conversations. Synthesised speech is used for services as diverse as the speaking clock, operator services, and reading e-mails over the phone.

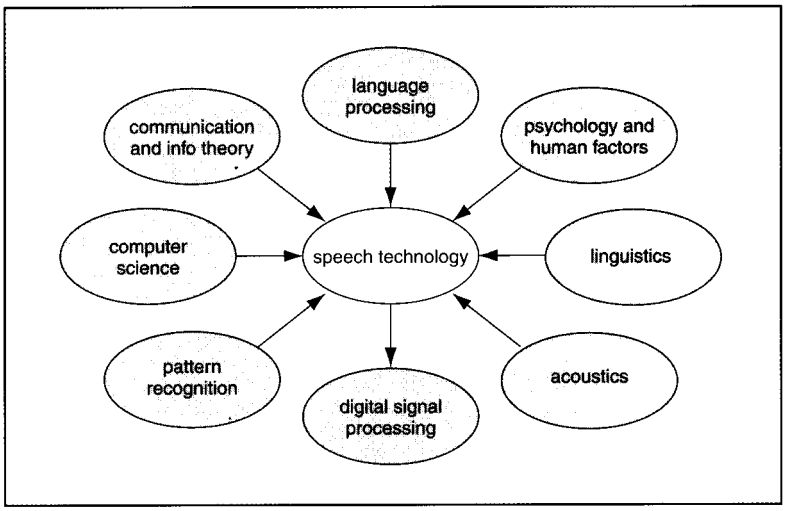

Fig. 1 Speech-processing technology disciplines

Speech technology is an amalgam of many disciplines (Fig. 1), and real-world speech applications increasingly require a broad-based, holistic approach to realise systems that are acceptable to the public at large. Engineering and computing need to be complemented by expertise in man–machine interaction, human perception, psychology, acoustics, linguistics, natural-language processing, and

*This paper is based on a paper originally published in the *BT Technology Journal*, January 1996, Vol. 14, No.1, pp.9–27

many other disciplines.

This paper provides a brief overview of the key enabling speech technologies and predicts some future trends and applications in this exciting and far-reaching field.

2 Applications and opportunities

Speech technology provides significant revenue opportunities through the introduction of new speech-processing services which can be easily accessed over the existing telephone, twenty-four hours per day, seven days per week. Speech technology can add value by:

- stimulating additional call revenues (for example, a network-based call-answering service which takes calls when the customer's line is busy or ringing with no reply)
- creating new speech services (for example, enabling in-flight telephone calls to and from aircraft and talking e-mail)
- providing differentiation in existing services (for example, speech enhancement and noise control)
- reducing operating costs (for example, by partial or full automation of operator services and call centres)
- allowing control of systems by voice to make them easier to use and hence improving customer satisfaction (for example, speech and speaker recognition)
- extending system usage to areas where other digit entry systems fail, where it would be dangerous to use hands, or where users require easily remembered access codes (such as names), and
- providing a simple means of getting large amounts of variable text data to users over the telephone (for example, by automatic conversion of text into speech).

Table 1: Interactive speech services

Voice store and forward
• answering service
Finance
• banking
• stocks and shares
• insurance quotations
• credit-card transactions
Entertainment
• betting
• horoscopes
• games
Information services
• timetables
• Yellow Pages
• news
Telemarketing
• promotions
Teleshopping/reservations
• theatres
• airlines
• catalogue shopping
Field operations
• data operation and retrieval
• field personnel job dispatch
• voice access to electronic mail
Automatic operator
• network services
• call centres

Developments in speech technology have enabled a new generation of interactive voice response (IVR) services operating over the telephone network. These range from Telco (telecommunication company) type services, such as automation of directory enquiries, to customer handling and information retrieval applications, which can offer commercial opportunities for major companies.

A major activity is the integration of speech technology with existing databases, IT processes and call-centre capabilities. These developments can range in scale from small bespoke systems involving single-line personal computer solutions to embedded network systems capable of supporting many hundreds or thousands of simultaneous calls. The larger-scale applications can significantly re-engineer the way in which customers interact with providers of information, goods and services.

Current telecommunications activities are primarily directed towards applying the technology to the telephone network for interactive speech services such as those shown in Table 1.

3 Interactive voice response systems

A typical interactive voice response system is shown in Fig. 2. *Speech input* is achieved by means of the recognition module (which may be supported by Touchtone™ or dial-pulse input), and *speech output* is by means of a synthesiser (which can be based on stored concatenated speech or text converted to speech). For a network-based application serving thousands of customers, a speech coder may be necessary in order to make efficient use of expensive storage media. The user aims to extract information from the system database (or voice store) by means of spoken commands.

Until recently, the performance of commercial speech recognisers restricted their use to relatively small vocabularies unsuited to applications where information such as names and addresses is required. However, the recent emergence of accurate large-vocabulary recognition at acceptable cost has increased the range of practical applications.

The *information manager* interfaces with the host application or database and acts as a mediator for high-level dialogue requests, for example building speech recognition vocabularies based upon the database content. The role of the information manager is described in more detail by Attwater *et al.* in Reference 2.

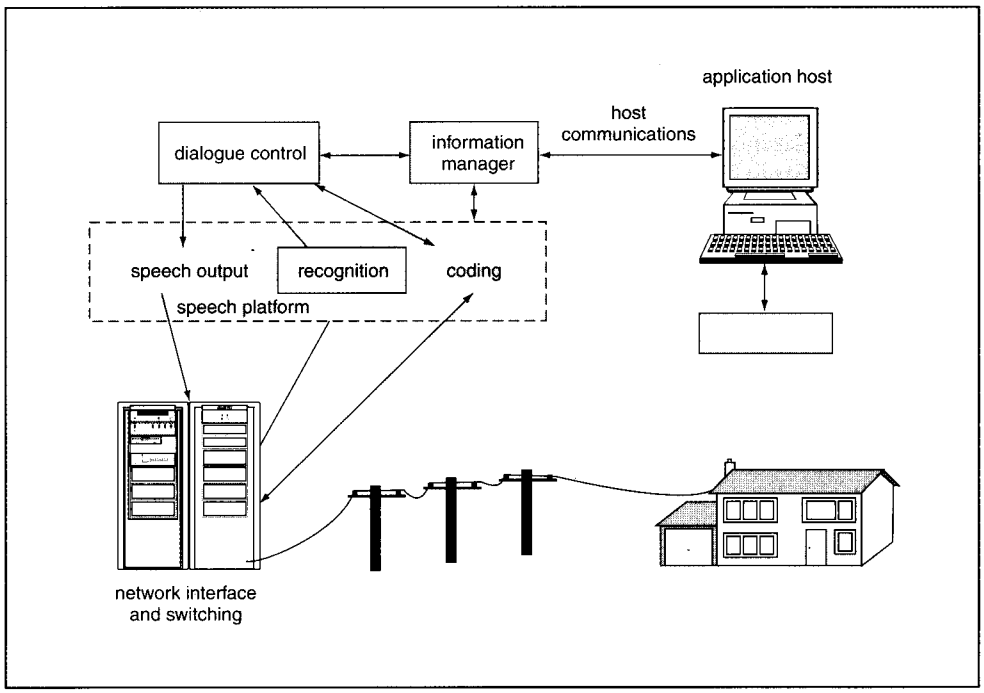

Fig. 2 Interactive voice response system

The *dialogue controller* ensures that the required information is obtained from the user in a controlled and structured way. It acts as a kind of gearbox matching the recogniser performance (engine) to the application (the wheels). Like any gearbox it can to some extent overcome deficiencies in the engine, but only if the engine is basically sound. Dialogues are not a substitute for inadequate or inappropriate speech technology.

Dialogue design

Dialogue design is often the key to successful IVR speech system design. It defines the system interface which the caller will encounter, guiding and interpreting each user interaction and selecting the appropriate response. The dialogue controller also ensures that the system recovers gracefully from any errors which are detected in the speech recognition, either by asking the user to repeat or by modifying the dialogue to obtain the required information by rephrasing the question.

From the user's point of view the system should be friendly, reliable and comfortable to use. To achieve this it is necessary to ensure that each component of the system performs adequately and aligns with the customer's expectations. It is a considerable challenge to orchestrate all the constituent technologies to best meet the expectations of the whole user population under all operating conditions.

Example of a dialogue

Within the UK, Directory Assistance (DA) is a chargeable service with a database of over 20 million residential, business and government entries involving nearly half a million distinct surnames, and nearly 30 000 distinct localities. An example of the dialogue for such an experimental application is given by Whittaker *et al.* in Reference 2. Partial spelling of the first N letters helps the recogniser distinguish between confusable names and homonyms (words which sound the same but which are spelt differently — typically 30% of names may be homonyms or synonyms of others).

In current IVR systems, the flow of the dialogue is controlled primarily by the system, not the user. Information is garnered one item at a time and a rigid dialogue structure enforces certain strategies for confirmation, rejection, error detection, correction and reprompting.

It can be frustrating for users to find themselves 'steered' through a complex dialogue with a machine only to find either that they are lost or that they have ended up with the wrong piece of information. This calls for a different approach to dialogue design, and systems are beginning to emerge where it is possible for the caller to respond much more spontaneously and flexibly and to take control of the interaction. For example, the next generation of interactive voice response systems might have the following 'conversation' with you:

How many messages do I have from Denis?

There are two messages from Denis Smith and one from Denis Crowe.

Can I hear the ones from Smith?

The first message from Denis Smith is about 'budgets'. It's 20 seconds long and a summary is '... (interruption) ....'

Give me the next one....

The next message is about 'the future of speech recognition'. It's 10 seconds long and a summary is ...

OK — um — read the previous one back to me in full.

This example illustrates a number of features which are

desirable in future interactive voice applications. The system does not need to prompt the caller constantly for explicit information. It uses discourse and domain knowledge to resolve ambiguous expressions such as 'give me the next one' and is able to understand references to earlier information. The caller can interrupt (barge in) at any stage, and the interrupting phrase is recognised. The caller controls whether the full message is heard (having been told beforehand how long it is), or whether a summary is given. The system discards paraverbals such as 'ums' but is able to identify words of significance wherever they occur (wordspotting).

Such a dialogue allows the initiative in the conversation to migrate from the computer to the user. More fluent interactions of this type will inevitably place greater demands on the underlying speech technology and require greater language knowledge than would be the case with more structured dialogues. However, spoken language will be an essential vehicle for searching and retrieving information and may help to make it accessible to all users, of all ages, and not just the computer literate. For example, cellphones are already too small to support a keyboard. For such applications, interactive dialogue-based systems will require both high-performance speech recognition and language understanding in order to identify properly and to respond to users' requests.

4 Speech technology

The mutability of speech

For many years speech was considered to be composed of linear sequences of elemental speech particles, or phonemes, put together rather like a string of beads. This misleading interpretation of the speech signal led engineers to imagine that machines could recognise speech simply by decoding the individual phonemes as they appear through time, in the same way as a modulated data signal could be produced.

In a similar vein, it appeared that synthetic speech could be produced by simply storing an example of each of the phonemes of a language and then sticking them together to form the new word. Unfortunately this proved not to be the case. Although it is convenient to visualise speech as a set of discrete symbols, in reality linguistic information is not encoded in discrete packets of time as imagined by this model, and traditional methods of dealing with coded signals will not work with speech. Our familiarity with text also encourages us to think of speech as 'atomic'. But, convenient as this picture is, speech is not like text:

> Therearenogapsbetweenwords. Sometim esthereareg aps inthe mid dleofwords. Acc*ents*, dialects and Stress *eggs*ist.

Compared with many other areas of digital signal processing, the challenges presented by speech processing are immense. Although the public imagination has been fired by images such as 'HAL' in Stanley Kubrick's '2001 — a Space Odyssey' the goal of producing complete natural language interfaces between humans and machines is still a long way off. The difficulty is due to the variability in factors that dictate performance:

- the great variety in the signal characteristics, when the word or phrase is uttered by different speakers and even when repeated by the same speaker
- the wide variety of characteristics of the channels through which the speech signals are sent and
- the nature of the accompanying background noise.

Above the level of the acoustic signal, ambiguities can occur at a higher level of linguistic abstraction. The phrase 'It's easy to recognise speech' could be interpreted as 'It's easy to wreck a nice beach' at a conference of surfers! To make matters worse, people do not always say what they mean or mean what they say.

This mutable behaviour is unavoidable and must be tolerated by a speech processing system which aims to provide a seamless and natural interface between people and machines.

The key underlying speech technologies are briefly reviewed next, starting with the most mature of the technologies — speech coding.

Speech coding

This is the process of converting speech into digital bit streams for efficient storage or transmission over band-limited channels. It can exploit redundancy in speech signals to reduce the transmitted bit-rate whilst at the same time exploiting the known properties of human speech perception to reduce the coding distortion to acceptable levels. Speech coders for particular applications are selected according to a trade-off between coding complexity, bit-rate and signal quality. The design trade-offs are further complicated by additional factors such as:

- codec delay (which makes echoes more audible)
- transparency (which affects the ability of the codec to pass non-speech signals)
- coder robustness to transmission errors and background noise
- tandeming ability (or the ability to mix different coders in a transmission path without the accumulation of distortion to unacceptable levels).

Speech coding has been subject to considerable international standardisation activities in recent years. Standards now exist for: 32 kbit/s (G.721); wideband (7 kHz) speech coding (G.722); 16 kbit/s (G.728); 8 kbit/s (G.729); variable-rate speech coding for circuit multiplication equipment (G.724) and for wideband packet networks.

Speech coders are in demand where there is a need to conserve radio spectrum. Examples include aeronautical telephony via satellite (Skyphone™ uses a 9·6 kbit/s speech coder), digital cellular (GSM Global Mobile System uses a 13 kbit/s coder) and digital cordless systems (CT2/DECT uses a 32 kbit/s coder).

They have also been applied in customers' private network applications, for example for speech-and-data multiplexers, in PSTN and ISDN videophone applications for the audio channel, and in interactive network-based voice messaging applications where cost of storage is still a critical factor.[3]

Coders are usually classified into the following three types:

- waveform coding
- vocoding
- hybrid coding.

The aim of waveform coders, as the name implies, is to reproduce the original waveform as accurately as possible. As these coders are not 'speech specific' they can deal with non-speech signals, such as background noise, music and multiple speakers, without difficulty. However, the cost of this fidelity is a relatively high bit-rate. Examples are pulse code modulation (PCM), adaptive differential PCM (ADPCM) and subband coding techniques.

In contrast, vocoders (voice+coders) make no attempt to reproduce the original waveform, but instead derive a set of parameters at the encoder which are transmitted and used to control a speech production model at the receiver. Typically, linear prediction coding (LPC) is used to derive the parameters of a time-varying digital filter, which models the dominant resonances of the vocal tract. Speech quality, although intelligible, tends to be synthetic and variable between speakers. Hence vocoding is not used for telephone network applications.

Hybrid coders combine features from both waveform coders and vocoders to provide good-quality, efficient speech coding. At rates between about 16 kbit/s and 4 kbit/s, good-quality coding is achieved by 'analysis by synthesis' techniques.[3]

Fig. 3 shows the performance of typical speech-coding systems as a function of bit rate. The quality scale is given in terms of 'mean opinion score' (MOS), which is obtained from informal subjective tests using different test material. The scale is from 1 to 5, interpreted as: 1=bad, 2=poor, 3=fair, 4=good, and 5=excellent. Telephony quality (toll) coders have an MOS rating of better than 4; good 'communication' quality codecs operate with an MOS between 4 and 3·5. As shown in Fig. 3, currently 'toll' or telephony quality coders operate down to bit rates of around 8 kbit/s, with the prospect in sight within a few years of rates down to 4 kbit/s and below.

In the longer term, more significant gains are theoretically possible if one considers that the maximum rate of articulatory movement is limited by the inertia of the physiological structures of tongue and jaw to about ten discrete sounds per second. Given the 44 phonemes of English, this gives a theoretical minimum of around 100 bit/s (even allowing for the inclusion of non-verbal cues such as emotion and phrase emphasis).

There is also considerable current interest in coding enhanced quality wideband speech (typically between 7 kHz and 20 kHz bandwidth) at rates of around 1 bit/sample for speech and 2 bit/sample for audio (including music), for ISDN applications such as teleconferencing, CD audio compression (MPEG-audio) and for commentary channel transmission.

Speech recognition

Today, when we interact with computers we normally have to resort to more artificial methods: keyboards, touchtone, touch-screens or mouse-driven menu systems. Speech recognition can improve the user interface by allowing spoken commands for accessing network services or for call routing, for example voice dialling over mobile phones.

The first paper on electronic speech recognition was published in 1952 (AT&T's 'Audrey' system) and described a system which could recognise single-digit utterances spoken in isolation by a single speaker. In the early 1980s BT began work on speech recognition which resulted in Topaz — the first repertory dialler for hands-free voice dialling in cars — in 1986. Early small-vocabulary speaker-independent recognisers were also deployed in the pioneering trial of automated banking in 1988, which at its peak involved some 400 Royal Bank of Scotland

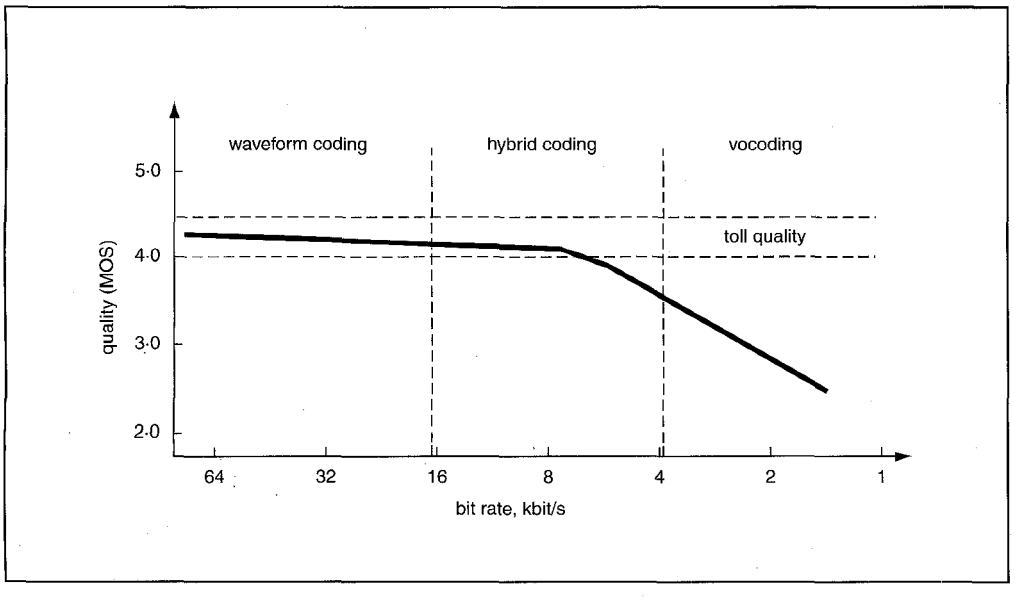

Fig. 3 State of the art in speech coding

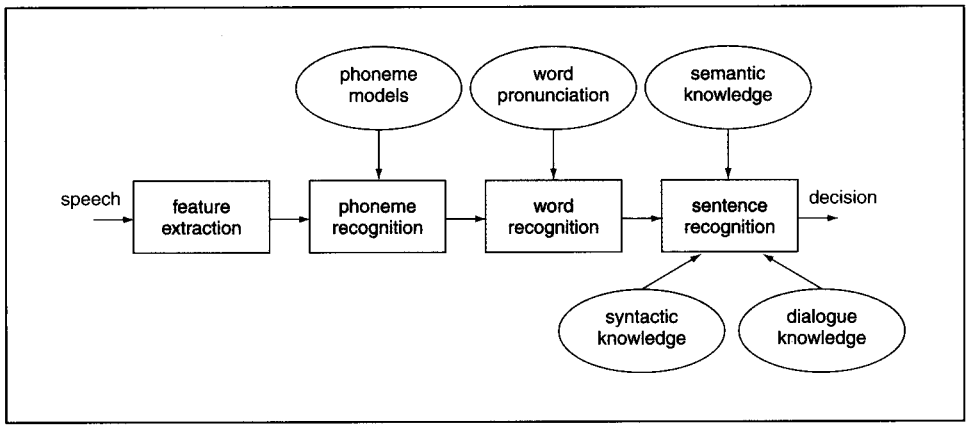

Fig. 4 Modern speech recognition system

employees in using an interactive banking system. More recently the latest large-vocabulary recognition has been implemented on the network-based Speech Application Platform.[2]

A speech recogniser is designed to recognise one of a set of words or phrases specified in a vocabulary. This is a difficult task, and not even a person is 100% accurate!

Fig. 4 shows the components of a speech recognition system. The front-end stage performs feature extraction, taking segments of the speech at regular intervals and transforming them to facilitate pattern recognition. Part of the process involves nonlinear filtering operations, which are assumed to occur in the inner ear. The resulting frames of features are then processed, either to identify words or parts of words (phonemes). In a more complex scenario, knowledge of the context can be used to aid sentence recognition by exploiting semantic (meaning), syntactic (grammar) and dialogue (discourse) constraints.

Many current speech recognisers use the technique of hidden Markov modelling (HMM) for pattern matching. A hidden Markov model is a type of model based on a statistical representation, and this helps the recogniser to cope with most of the variability in the way people speak. The recogniser aims to identify, from observing a sequence of speech features, which of the stored Markov models is most likely to have produced these observations. A good 'layman' description of speech recognition techniques, including HMMs, is provided in References 4 and 5.

A *connected-word* recogniser has to match utterances to strings rather than to single words. This is usually achieved with a finite-state network of word models representing the whole vocabulary. The shortest path through this network which matches the utterance is selected as the recognised string.

Recognition performance depends on the vocabulary to be recognised, the confusability of words within the vocabulary and the number of users. A PC-based business dictation system has a large vocabulary (typically 20 000 to 30 000 words), but must be trained before use for a particular speaker and is able to exploit tight constraints on language and word statistics. The speaker is also required to adopt a disciplined approach, speaking clearly with short gaps between words.

In contrast, speech recognition for telephony is likely to use only a small vocabulary, since the recognition task is far more challenging, with much greater signal, network and speaker variability. Also context and grammar constraints tend to be weaker, as with surname and address recognition. The first trial of a directory enquiries system in which both large vocabulary and spelt input were used was undertaken in East Suffolk in 1994. This system worked for a directory of 25 000 names. Today a similar system is available at BT Laboratories giving access to the telephone numbers of everybody on the site — 5000 people.

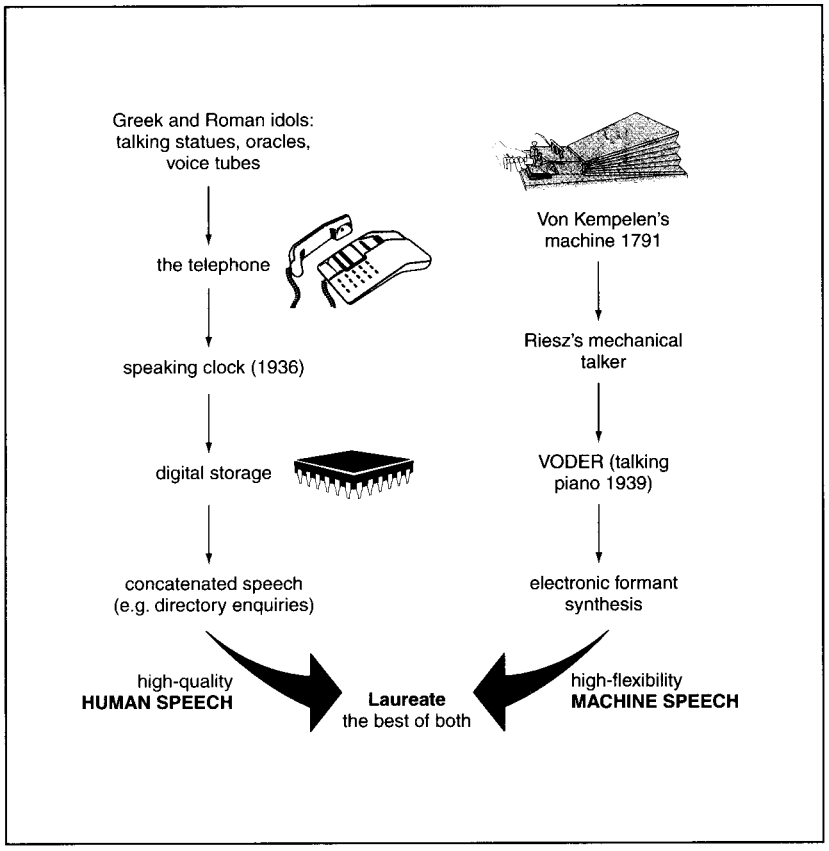

Fig. 5 Evolution of speech synthesis

In the past, the production of new vocabularies took several months and required many examples of the words from a hundred or more speakers. Also the whole process had to be repeated if a change or addition to the vocabulary was subsequently needed. Recently, a new method has been developed to allow vocabularies to be generated very quickly by simply typing the words to be recognised. Word models may be built from phonetic transcriptions using either a dictionary or a set of letter-to-sound rules. Because of the various approximations, recognition accuracy for text-generated vocabularies is generally inferior to that for systems which have been trained on actual speech utterances. Furthermore, in applications where the vocabulary is changing rapidly, such as directory or customer identification applications, the system must be able to update dynamically its active vocabularies.

Speaker-dependent repertory dialling (e.g. 'Call Bob') has been available for some time. Typically, this requires the user to repeat the word or phrase a small number of times, under the guidance of the dialogue controller. Recently, this concept has been extended to speaker-independent operation.

Speaker recognition

Speaker *verification* technology can be used to validate the claimed identity of a person from his or her voice-print. It is often combined with a general-purpose connected-speech recognition system to recognise a spoken personal identification number (PIN), and then to check that the PIN was spoken by the authorised person.

Speaker *identification* can be used to recognise a member of a closed user group (for example a family member) directly from a known spoken word (text dependent) or from any spoken utterance (text independent).

Current applications include secure access control to information, banking, computer networks, PBXs and work areas. The technology can also be used to provide access to a range of network services according to customer profiles, such as name dialling and travel information.

Speech synthesis

The evolution of speech synthesis technology is illustrated in Fig. 5.

In 1791 Von Kempelen constructed a talking machine (Fig. 6), which consisted of a bellows, a mouth shape, nostrils and whistles. The machine included a compressible leather tube and an air chamber equipped with a reed leading to a soft leather resonator which could be manually shaped for the formation of the vowel sounds. Consonants were created by holes which the 'player' closed by movement of the fingers. The Von Kempelen machine could produce about twenty different sounds!

The earliest electronic synthesis of speech was achieved by Dudley in 1939 when he demonstrated a manually controlled speech synthesiser known as VODER (Voice Operated Demonstrator) at the New York World Fair. After the Second World War, development began in the Post Office of techniques to analyse and synthesise natural speech. This led to a variety of synthesis systems based on the source-filter model of speech production, of which the formant synthesisers developed by the Joint Speech Research Unit in the UK and MIT are perhaps the best examples.

Speech output from computer-based equipment has been commonly achieved by generating messages from stored speech fragments (as with the BT speaking clock). These messages must be recorded by a speaker and, although a natural-sounding voice output is achieved, significant amounts of data storage are required. An additional constraint on system design and extension arises when the messages need to be changed or updated if the original speaker is not available. There are many applications where the versatility of a full text-to-speech system is the only practical solution.[1,6]

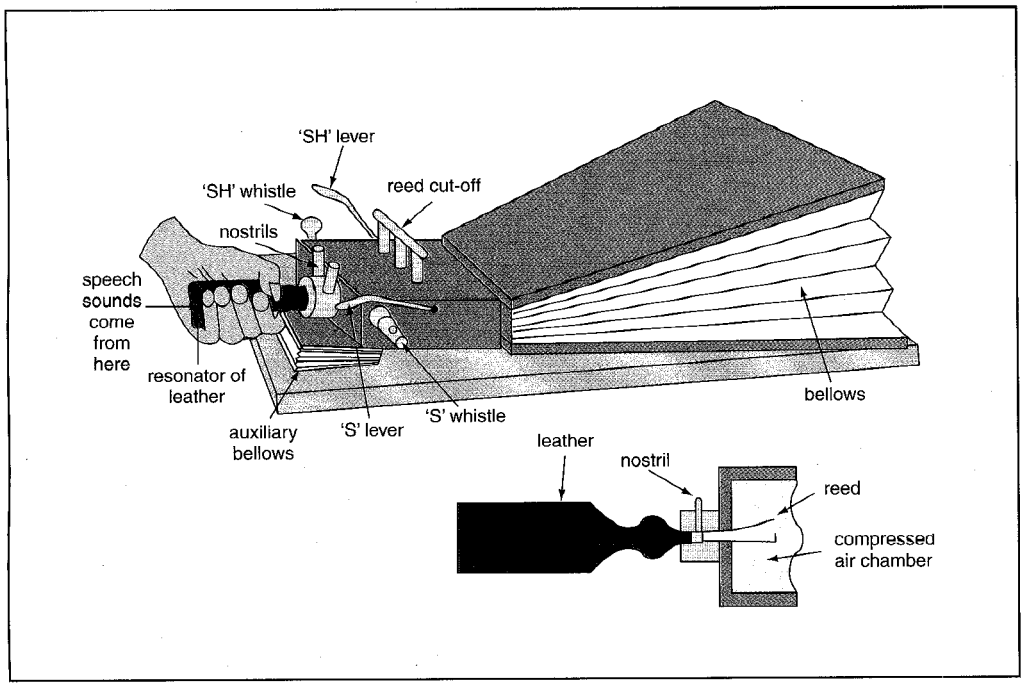

Fig. 6 Wheatstone's replica of Von Kempelen's machine

A text-to-speech conversion system, such as BT's Laureate system, is designed to convert unrestricted textual input into speech. The latest synthesisers are an improvement on the earlier speech synthesis systems in that they do not attempt to generate sounds artificially to mimic human utterances. Rather, the speech is constructed from elemental components of a person's recorded voice. This preserves the voice characteristics of the original speaker and hence the system can be easily modified to effect local accents and other languages.

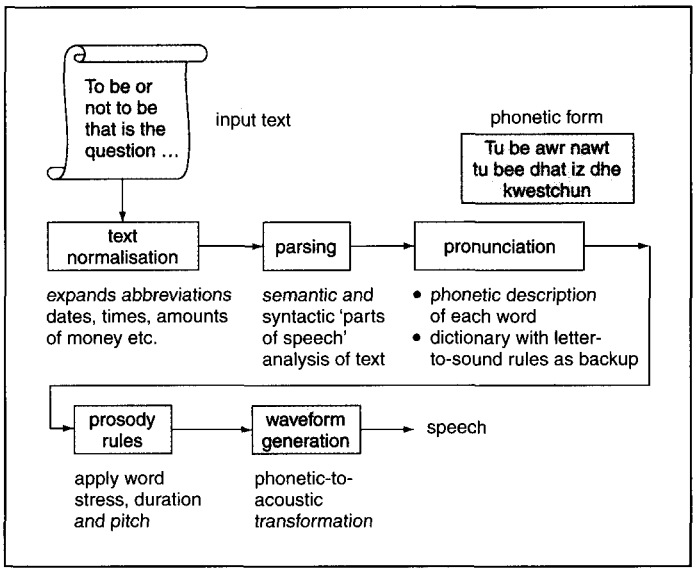

Fig. 7 Text-to-speech system

A typical text-to-speech system is shown in Fig. 7. The first stage of processing involves expanding the input text by a *normalisation* stage which deals with abbreviations and acronyms, such as 'St. John St.', which is expanded to 'Saint John Street'. This is followed by a stage of syntactic parsing, which aims to resolve ambiguities between words such as 'lives' or 'convict' which have have both a verb and a noun form. A semantic analysis may also be carried out which provides a representation of the meaning of text as an aid to *word pronunciation*. Next a phonological analysis of the words is made to derive an appropriate word pronunciation, based on a large dictionary and a set of letter-to-sound rules, where the dictionary look-up fails (e.g. for proper nouns). *Prosody rules* are then applied to provide the required emphasis on duration, intonation and major word stress. Finally the speech is *synthesised* by looking up the appropriate speech segments, and concatenating them in as seamless a fashion as possible.

An example of an application for text-to-speech synthesis is a telephone-based catalogue ordering service, where the system can respond with a full description of the item as well as saying the name and address of where the item is to be sent. Here the quantity of information would be too large to record directly for the average catalogue. An example of a system where the output information must change rapidly is access to news. Here large amounts of text can be updated regularly and made available over the telephone with no manual intervention. The ability to speak e-mail messages over the telephone is another important application, enabling people to check their messages when away from a computer terminal.

Speech enhancement

This is the process of enhancing the perceived quality of a speech signal received over a telephone circuit. This area covers a wide range of applications, such as noise and echo control, including the cancellation of echoes on long international circuits to enable simultaneous two-way speech communications. Echoes occur when signals undergo delayed reflections and can affect both talkers and listeners on a call. The effects of telephony echo on customers are complex, depending on prevailing electrical and acoustic factors, as well as the motivation of both customers.[7]

Echo suppressors have been used for many years and conceal echo by detecting when the distant customer is speaking and the near customer is silent.

The alternative of echo cancellation was first proposed by Sondhi of AT&T Bell Labs, and the first single-chip custom VLSI implementation appeared in 1969. As shown in Fig. 8, a replica of the echo signal is synthesised by an adaptive filter which models the echo path and is subtracted from the send signal. In this normal mode of use

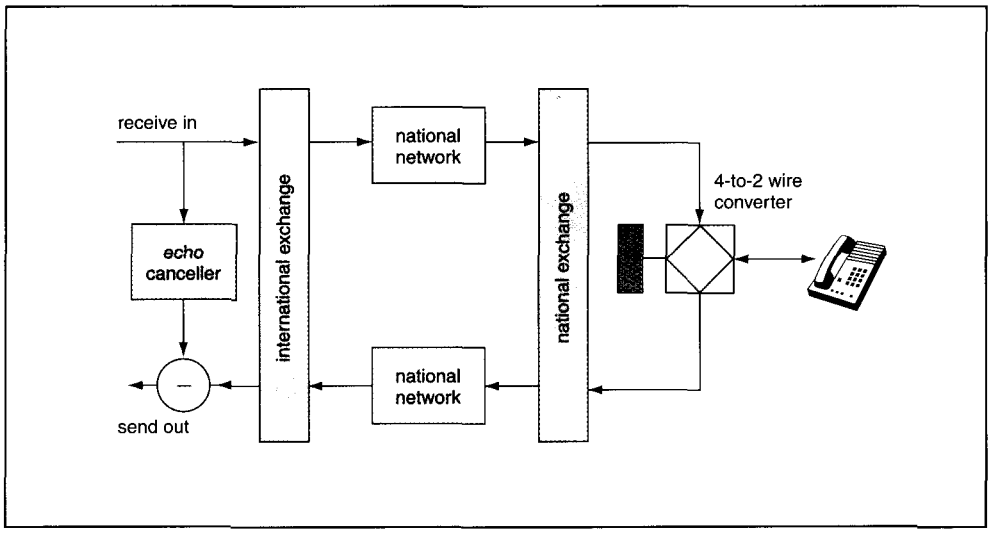

Fig. 8 Location of network-based echo cancellers

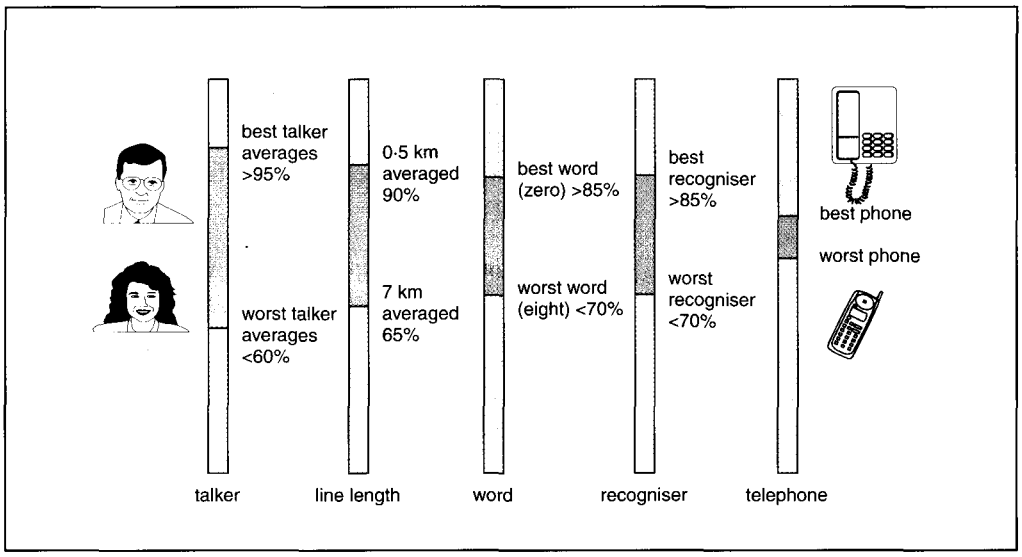

Fig. 9 Factors affecting recogniser performance

the echo canceller is only required to model the relatively short duration echoes returned by the national network. In general it is desirable to cancel echoes as close to their source as possible. Cancellers offer somewhat better quality than suppressors on long-delay circuits, because they have greater signal transparency and less intrusive speech-clipping effects.

Interest has also grown in applications for adaptive noise suppression, where the speech and noise typically have overlapping spectra.[8] In such a system adequate performance is dependent on obtaining a good estimate of the noise spectrum, either from a separate reference microphone or from noise-only periods in the incoming speech obtained using a voice-activity detector.

Applications include very noisy environments where the signal-to-noise ratio is low, such as telephone kiosks in busy concourses or in financial dealer rooms.

Signal processing is also employed in speeding-up (or slowing-down) the playback of recorded messages (without introducing noticeable distortion to the speech), and to eliminate speech-clipping and other nonlinear distortions that are occasionally encountered in public switched telephone networks.

5 Speech technology assessment

Depending on the system being considered, an assessment of subjective performance may be made by means of either a conversation or a listening test. In a listening test, people (who are called subjects) are placed alone in a controlled environment and asked to listen to speech transmitted through a series of randomly presented network connections (which are called conditions). At the end of each condition the subject is asked to vote on their opinion of the speech heard. A 5-point mean-opinion-score scale is used, as indicated in Fig. 3. For example, a speech coder suitable for switched telephone network operation would have a mean-opinion-score rating of above 4·0.

During a conversation test two subjects are placed in separate rooms and asked to converse over randomly presented conditions. As an aid to conversation they will be given a simple task to perform. At the end of each condition they each vote on the condition without discussing it with each other.

In some cases where the requirements of the final system are not well understood or if a suitably accurate prototype of the recognition subsystem is not available, 'Wizard-of-Oz' experiments are sometimes conducted with a trained human operator in place of the speech recogniser. This allows the focus of the experiment to be maintained on the aspects of dialogue interaction of interest without the complication of errors introduced by the recogniser.

Comparative objective tests are conducted on in-house and commercial speech systems. Fig. 9 shows the results of such an assessment on several commercial speech recognisers (averaged results). As can be seen, the talker variability has the greatest effect, with the telephone instrument contributing least to recognition errors.

Recent studies on human perception of speech which has been subjected to nonlinear network distortions have been encouraging, with a high degree of correlation between the derived objective measures and people's subjective opinion of the impairment, as indicated by Hollier *et al.* in Reference 1. There is research interest in extending this modelling approach to joint perception of mixed-mode (video and speech) signals for emerging multimedia terminals.

6 Future directions in speech technology

Interactive voice systems

An essential ingredient in the success of any business is access to information sources and modern telecommunications can give direct access anywhere in the world within seconds or minutes. With the spread of personal communication products, many of them too small to support a keyboard, speech and language processing, incorporating some degree of understanding, will play a necessary role in providing access to, transforming, and delivering information from these vast resources.

As interactive speech systems are deployed operationally they will generate new field experience and customers' service perception data. Such 'real-world' data

will allow a new R&D focus on technology improvement. Advanced data visualisation techniques will also help researchers to interpret this volume of data. By examining large numbers of interactions, statistical models of dialogues can be generated and used to optimise deployed dialogues developed using traditional techniques. This offers the potential for a unified understanding from low-level signal processing to higher-level semantics.

Speech coding

Optical-fibre systems provide virtually limitless bandwidth in terrestrial telecommunication networks, so the need for speech coding may diminish on circuit-switched point-to-point connections. However, with the growth in corporate 'virtual' networking and wideband packet networks, the added flexibility afforded by variable-rate speech coding will be in demand for competitive managed-bandwidth schemes.

Bandwidth will also remain under pressure in mobile and personal communication systems. In such areas, the demand for speech coding will continue to grow strongly over the next decade. This demand will be driven by rising customer expectations on quality, by international standardisation, and by technical advances in the on-chip integration of digital and analogue functionality.

There is also considerable interest in using coding techniques to enhance the speech transmitted over networks and for CD-audio compression for future music and multimedia distribution systems.

Speech synthesis

The current drive to improve the naturalness of text-to-speech systems will continue with the emphasis in areas such as:

- improving the 'natural rhythm' of synthesised speech (phonotactic variation)
- providing more choice in selecting the 'personality' of the speaker (current systems tend to adopt a single speaking style)
- providing more choice of language and accent with the associated tools to provide such flexibility in text-to-speech systems.

As full-motion video requires much bandwidth and/or suffers processing delays, there is growing interest in synthetic model-based image processing coupled with text-to-speech systems. In such a system, a wire frame model of the person's head is created from photographs and a synthesised voice synchronised to the lips of the model (Fig. 10). By typing in text the synthetic persona can be made to 'speak'. Such technology shows promise for new services such as customised database access systems and 'continuous presence' conferencing and to provide a research framework for studying fundamental mixed-mode signal interactions.

Speech recognition

Undoubtedly the current emphasis on accuracy and processing efficiency for large-vocabulary, speaker-independent recognition will continue. To make such systems scalable to national network applications, more work will be needed on improving noise robustness and out-of-vocabulary rejection. There is also interest in developing systems that can adapt to the characteristics of the user's voice and the transmission environment. Techniques such as 'barge-in' and 'wordspotting' will also grow in importance in the drive to make speech systems more natural, and hence more acceptable.

To enhance the performance of recognisers, significantly more basic knowledge will be needed in areas such as feature extraction, and building-in of more knowledge of the physiology of the ear and brain as a

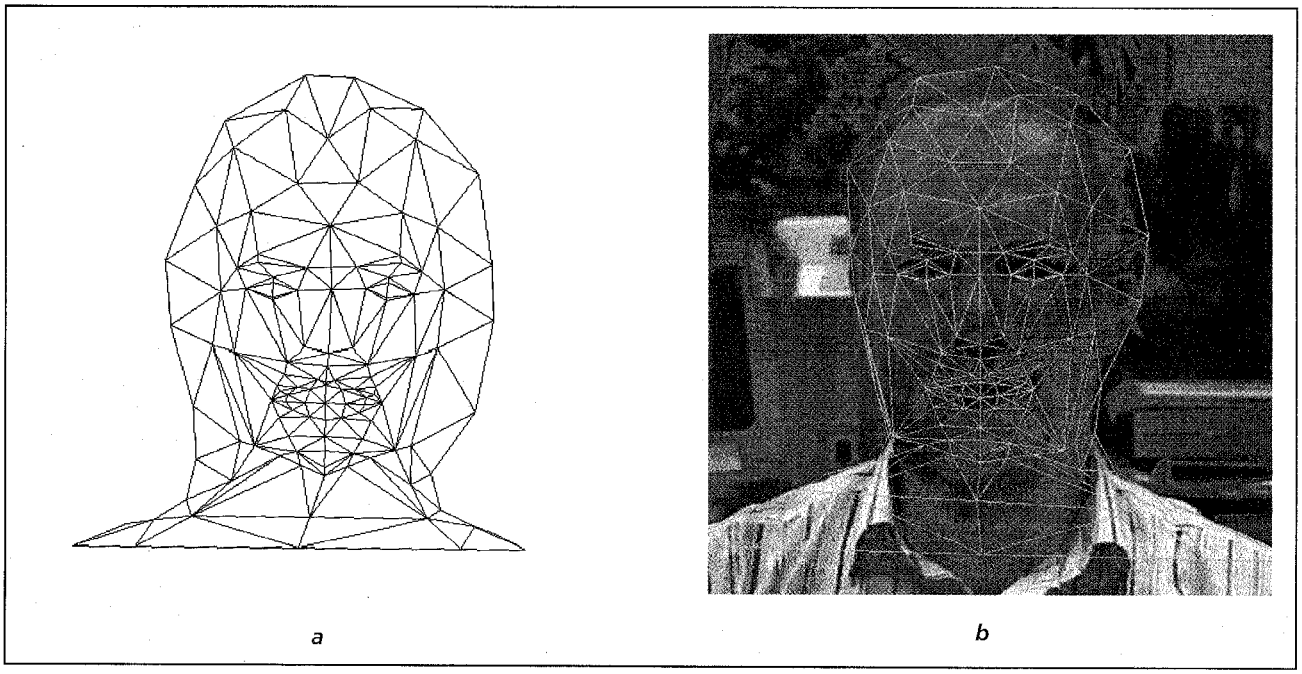

Fig. 10 Synthetic persona (talking head). A wire frame model (*a*) is derived from a photograph (*b*)

coupled 'system' as it becomes available. Improvements to modelling of speech can be expected in a drive to reduce the limitations of the existing statistical-model-based approaches. New paradigms based on mixed-mode (acoustic, speech, visual and gesture) cues may offer one way forward.

There will also be an increasing interest in multilingual recognition and the associated ability to quickly recognise the language of the speaker. Ultimately, the ability to identify the topic domain as a means of focusing the recognition resources will be developed.

Person-to-person: the future of telephony

Much of today's speech technology is concerned with man-to-machine interfaces using voice. But what about the future of person-to-person communications?

Today, videoconferencing is limited to viewing the participants through a small rectangular window. Images are captured through a single camera and sound by means of a single microphone. In 'hands-and-eyes-free' mode the speech often sounds clipped or as if it originated in a bathroom.

There is interest in acoustic-echo cancellation, adaptive beam-stearing (and talker location) and intelligent loudspeaker technology for 'last metre' processing to remove reverberation and noise in hands-free conferencing systems — to improve the naturalness of such interactions.

Early work in this area at BT Laboratories has shown the potential of this technology to provide improved naturalness and utility for loudspeaking telephones.

In the research environment teleconferencing can be made very realistic, as with electronic work spaces with eye-to-eye video conferencing that maintain gaze awareness. When augmented by directional and speaker-tracking audio reproduction the illusion of 'being there' can be almost complete.

Speech and language processing has the potential to make such interfaces more natural and easy to use through the exploitation of visual, acoustic and gesture cues. Herein lies the true destiny of speech processing — not just one mode of communications, i.e. speech, but all the senses orchestrated to meet the real underlying communication needs of people.

7 Conclusions

Speech communication is, and will continue to be, key to the use of BT's network, currently accounting for well over 90% of revenue. Speech is the most natural way for users to communicate from person-to-person and in the future, where appropriate, from person-to-system.

Despite the enormous effort expended on speech technology research over the last few decades, still surprisingly little is known about signals — or rather about how humans perceive and interact using them. For example, in areas such as speech recognition, much more fundamental knowledge is needed to allow machines to be made that approach the capability of humans. Arguably, there has been too much recent emphasis in the digital signal processing, or 'DSP', community on solving the problems of *digital processing* and not enough on understanding the subtlety of *signals* and human perception.

As speech technologies are commercially exploited different skills and expertise will be required at different stages in their evolution, from research/prototyping in the early stages to downstreaming and support of real-world applications in the later stages. The effective coupling of research to delivery of applications is critical. In particular, the problem of scalability, from laboratory demonstrator to national network service is fundamental.

Speech processing is a dynamic, exciting and commercially relevant field, overlapping the traditional subjects of mathematics, electronics and computer science. Speech technology is well poised to affect dramatically the way we all communicate and interact in the 21st century.

Fred Westall received a BSc(Eng) degree in Electrical and Electronic Engineering from University College London and an MSc degree in Communication Engineering from UMIST in 1973 and 1975, respectively. Following graduate training and a spell as a microwave development engineer he joined BT Laboratories in 1975 to undertake research and development of speech-band modems for the PSTN. He has been closely associated with digital signal processing ever since. In 1982 he became head of the Speech Coding Applications Group and in 1986 was appointed to manage the Data Products Development Section, where he was responsible for packet terminals development and high-speed modem R&D. His most recent responsibilities at BT were for downstreaming speech-band applications onto DSP speech platforms and for signal processing R&D, notably in the fields of speech recognition, coding, analysis and synthesis. He is now with Brite Voice Systems. He is an IEE Fellow.

Address: Brite Voice Systems, Brannan House, 4 The Cambridge Business Park, Milton Road, Cambridge CB4 4WT, UK.

References

1 *BT Technol. J.*, Theme Issue on 'Speech technology for telecommunications', January 1996, **14**, (1)

2 *BT Technol. J.*, Theme Issue on 'iSAP and its applications', April 1996, **14**, (2)

3 Boyd, I.: 'Speech coding for telecommunications', *Electron. & Commun. Eng. J.*, October 1992, **4**, (5), pp.273–283

4 Johnston, R. D.: 'Speech recognition for speech services', *BT Eng. J.*, July 1994, **13**, Pt. 2, pp.145–158

5 Talintyre, J. E.: 'The listening phone', *IEE Review*, July 1996, **42**, (4), pp.151–154

6 Breen, A. P.: 'Speech synthesis models: a review', *Electron. & Commun. Eng. J.*, February 1992, **4**, (1), pp.19–31

7 Lewis, A. V.: 'Adaptive filtering — applications in telephony', *BT Technol. J.*, January 1992, **10** (1), pp.49–63

8 Vaseghi, S. V.: 'Advanced signal processing and digital noise reduction' (Wiley Teubner, 1996), ISBN 0-471-95875-1

Received in final form 16th September 1996