Proceedings of the Institute of Acoustics

THE FUTURE OF SPEECH TECHNOLOGY

Michael J. Carey

Ensigma Ltd, Turing House, Station Road, Chepstow, Mons, NP6 5PB, U.K.
michael@ensigma.com

1. INTRODUCTION

The author believes that the development of speech technology has reached a critical point. Many workers in the field appear discouraged by a perceived lack of progress in fundamental research even though the commercial applications of speech technology are becoming ubiquitous. The cause of this disquiet is that the focus of the subject is moving towards applications which require a different technical emphasis. The technological environment in which speech technology will be deployed is changing rapidly. To use speech technology successfully these changes must be anticipated so that the speech interface becomes a worthwhile part of the system in which it is used. While speech research has provided many interesting intellectual challenges over the last fifty years and is a worthwhile scientific pursuit in its own right speech technology will become an irrelevant technical backwater if it is not properly integrated into the target system in a way that users find acceptable. It is only one of several competing technologies for the human-machine interface. While it may be the most effective form of communication between people it is not necessarily the best between people and machines. It is therefore important to identify the engineering context in which speech technology will be used and to direct our efforts towards making the technology useful for the user.

In this paper I attempt to address how we, the speech research community, might approach the future challenge of making speech technology of real benefit for people. To be that they must become useful and widespread and so these are the criteria by which I will judge the success of applications. I have deliberately ignored special applications such as the military market which are important in their own right and where much progress has been made but outside the experience of most people. In the remainder of this paper Section Two briefly reviews the state of the art in each main area of speech technology. Then in Section Three key developments in consumer electronics, communications and semiconductor technology are reviewed and conclusions about the future environment for the application of speech technology are drawn. In Section Four consideration is given to the important areas and applications for speech technology which will result from the developments described in Section Three.

2. THE STATE OF THE ART

This brief review of the state of speech technology deliberately limits itself to systems which are in use, not just in the laboratory. For many years there have been too many inflated claims about products just about to be launched to give much credence to anything which is not available in the market place. In addition a sustained presence in the market place is one of the few signs of user acceptance.

2.1 Speech Coding

Speech coding has to be regarded as the outstanding success of speech technology. While other applications of speech coding exist such as in private networks and voice mail this success is mainly due to the widespread use of digital mobile phones. Estimates vary but 500 million are expected to be in use in 2002. Since each phone will have between one and six speech coders there will be more than one billion speech coders in use at that time. This will dwarf numerically all other applications of speech technology.

Date	Standard	Rate	Technique
1985	CCITT G721	32kb/s	ADPCM
1989	CCITT G728	16kb/s	Backward Prediction with Vector Quantisation
1993	CCITT G729	8kb/s	Code Excited Linear Prediction
1999	CCITT G7??	4kb/s	Prototype Waveform Coding?

Table 1 Toll-Quality Speech-Coding Standards[1]

Why has this come about? Speech coders are a real solution to a real need. Since the adoption on coders based on linear prediction the bit-rate of toll quality[1] speech coders for telecommunications transmission applications has halved at five year intervals and is now approaching 4kb/s. These coders or there variants have been included as part of mobile system standards such as GSM. The scarcity of available spectrum is an important problem in the growth of digital cellular communications. Low bit-rate speech coders help alleviate this and the growing use of mobile phones in buildings is helped by error tolerant speech coders reducing the number and power of the base stations required.

[1] Toll Quality means a coder with a Mean Opinion Score of about 4.0(good), e.g. 64kb/s pcm scores 4.1

2.2 Speech Synthesis

In comparison with speech coders the speech synthesiser has had more mixed success. It is now twenty years since the launch of the Texas Instruments 'Speak and Spell' and the Kurtweil reading machine. Since that time speech coders have increasingly encroached into the applications in which synthesis was used. Synthesis was preferred to coding for two reasons, the lower effective bit-rate and the ability to output arbitrary messages. Against this was the clear superiority in the quality of coded speech over synthesised speech. With the progress in reducing speech coder bit-rate described above the first advantage of synthesisers has disappeared. A 1Mbyte memory can store more than half an hour of speech coded at 4kb/s. The future for the large scale use of synthesisers would appear to lie in their ability to output arbitrary text such as reading email over the phone.

2.3 Speech Recognition

The automatic dictation machine which translates arbitrary speech into text was a distant goal only twenty years ago. That a 60k word continuous recogniser should be available for less than £1,000 including a multi-purpose computer would have been incredible then. There are now several millions of these systems in use[2].

The last year or so has seen the introduction of the repertory diallers built into hand-held mobile telephones. While the systems are small vocabulary and speaker dependent they incorporate algorithms such as non-linear spectral subtraction to improve the robustness of the isolated word or phrase recognition. Given the rapidly growing market for mobile telephones it would only require a small percentage of these phones to have recognition for this to rapidly become the most widespread application of speech recognition.

Some small vocabulary speaker independent recognition such as the BT Call Minder system are in widespread use. In North America the applications of network based recognisers is more widespread. It is claimed that ATT saves $400M each year from the automation of simple operator services using a comparatively simple word spotting system . The present generation of dialogue systems using continuous speech recognition is close to commercial launch. Systems for railway time-table enquiry and magazine subscription have demonstrated impressive performance. However the degree of long term user acceptance is yet to be established.

2.4 Speaker, Language and Gender Identification.

Language and Gender Identification are essentially solved problems and in use for forensic purposes. Also the performance of the present generation of text independent speaker verification systems is good enough to be useful. It may be therefore surprising that text dependent which is inherently more accurate speaker identification has not been accepted for access control applications. The key problem appears to be that people do not want to be bothered with security systems unless they are speedy, unobtrusive and faultless. Unfortunately speaker verification is not perceived as quick or unobtrusive and is certainly not faultless.

3. THE EVOLVING ENVIRONMENT

Twenty years ago the idea that a large vocabulary recogniser would run as software on a cheap personal computer was unbelievable. Then the proposed solution was a centralised server accessed via the telephone network. While many servers have been employed the number used has been dwarfed by those realised as PC applications. Similarly the chief application of speech coders was expected to be in long distance fixed links such submarine cables or in private telecommunications networks. The chief application has been in short distance wireless links in digital mobile phones.

To evaluate the future of speech technology it is important to predict the environment in which it will operate, that is to understand the future growth of the technology we need to look at current developments in the computer and telecommunications industry which will not be a simple extrapolation of current trends and then to try and predict how the technology will be shaped by these developments.

3.1 The Internet

The growth of the Internet continues to accelerate. Internet traffic in the US grew 100% in a three month period last year. In the San Francisco bay area data traffic now exceed voice traffic. In the U.K. about a fifth of people have email and 40% of companies are connected to the Internet. While the mechanism for internet access, Personal Computer, Web TV, or personal organiser may be disputed the requirement for access is indisputable [3].

3.2 Mobile Communications

The growth of the internet is mirrored by the growth in mobile communications. Here however the leading edge is in Scandinavia where there are ten new mobile numbers for each new wired telephone connection [3]. As we saw in Section 2.1 the total number of mobile phones is set to exceed 600m by 2002. It is also anticipated that 10-15%of these (that is 60 - 90 M phones) will be smart phones such as the Nokia Communicator or the Alcatel One Touch [4] providing internet access including email and Web browsing in addition to voice.

3.3 Semiconductor Technology

The changes described above have been brought about by the relentless progress of semiconductor technology. Current designs use a feature size, effective line width, of 0.25 micron. New designs will be implemented in 0.13micron giving a fourfold increase in the component density. It will be possible to fit 223M transistors, 26M gates on a 20 by 20mm device [5]. With careful low-power design this raises the possibility of small handheld devices with the computer power of the present generation of PCs but running on a couple of batteries.

3.4 The Future Scene

The PC will continue to be a growing area of application for speech technology. However few people want a PC per se. It is the PC's functions, spread sheet, database word processor etc, which are wanted. PCs will only be desired while they are the most effective way of fulfilling these functions. The office will probably require PC functions for the foreseeable future. However some of the PC market will be replaced by more specialist and mobile devices. The car environment with its increasing use of electronics for entertainment, navigation and communication as exemplified by the VODIS [6] project is likely to be an important area for speech input output applications since the drivers hands and eyes are occupied.

The developments in the internet wireless communications and semiconductors lead to the possibility of a new class of portable devices, Wireless Information Devices, which will combine the functions of mobile phone, organiser and internet terminal. These devices are the natural place for the deployment of speech technology and are likely to be the main growth area in telecommunications. The present generation of organiser struggles to incorporate the keyboard in a reasonable size. Some like the Apple Newton or the very successful 3Comm Palm Pilot [7] have dispensed with it

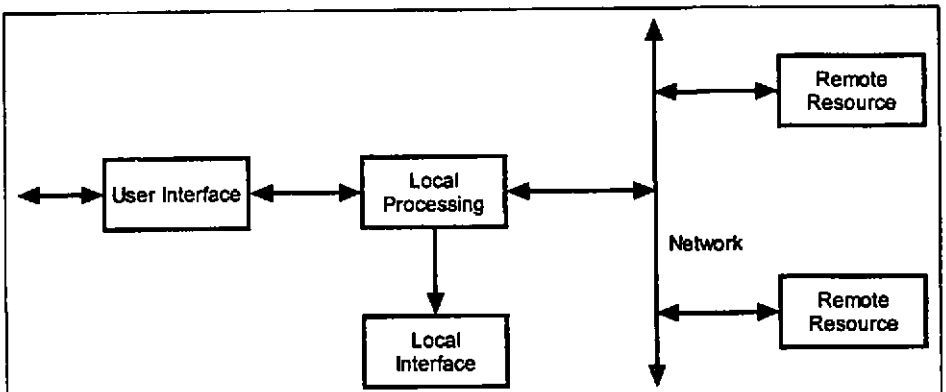

Figure 1 A General Information Access Architecture.

substituting hand writing recognition of dubious utility. The inclusion of speech technology would transform the user interface in these devices.

Home applications of speech technology have been discussed for many years. The control of intelligent appliances has been suggested on countless occasions. It now seems that the bluetooth piconet standard [8] will ease the interconnection of domestic devices allowing them to be controlled from a mobile bi-directional controller.

The most important aspect of these devices whether they are PCs, car systems, Wireless Information Devices or Home Controllers is that they will be networked and multi-modal including several different forms of input and output device. They will all share a common architecture as shown in Figure 1. This may seem obvious but it puts the user interface locally. Also speech will be one of the options available for the user interface but it will only continue to be available in later generations of devices if it is used.

4. THE IMPLICATIONS FOR SPEECH RESEARCH

Stressing again that speech technology is mainly about the user interface there seems to be every reason to keep this interface and hence the speech technology as close to the user as possible. Then user specific information can be stored with and move with the user, the idea of *locality*. Locality confers several benefits the chief being it makes the interface independent of the network and the service provider allowing the user to roam or change service providers without the loss of user specific information. The

penalty for this is that the savings in hardware resulting from the shared use of servers on the network are lost. However as we have seen in the previous section the computing capability of a single silicon device is potentially so large that such savings are likely to be negligible.

These systems will also inter-work with the *intelligent agents*. The recognition process can be greatly reduced in complexity if the system is able to anticipate to some degree the users requirements. Intelligent agents which learn about the user will interact with the recogniser to simplify the process.

Dialogue systems are presently designed for the voice network assuming that the communication will be speech input and output. Much of the dialogue is concerned with reducing the information flow to the user so that the output can occur in a reasonable time and the user is not overwhelmed and is able to remember the useful parts. In future this need not be the case. If most terminals incorporate a screen more information can be output in the same time and this information can be stored and displayed locally. The system then needs far less interaction with the user as the later stages of information selection can be carried out by the user visually not by the machine as a result of further dialogue. Visual output can also provide instant feedback of the results of the input recognition process leading to improved strategies for error recovery. This capability for *multi-modal* operation will be essential in future systems.

Robustness will be an important feature of future systems. The need for recognisers to perform well in channels with added noise or frequency distortion should reduce with the positioning of the speech interface close to the processor. However the problems cause by background noise including other speech and music or reverberation will increase as the technology is more widely used and must be addressed.

Speech coding research may soon become a victim of its own success. Is there any real need for bit rates below 4kb/s? Further research should centre on higher bandwidth coders especially since the imperative of 8kb/s dictated by the present digital network will weaken as the digitisation and coding functions are located in the handset and UMTS delivers even higher bit-rates. Another useful improvement would again be better performance in the presence of background noise with perceptual noise reduction techniques included into the coder. The applications of speech coding will continue to grow with more emphasis on its use in internet telephony and voice mails over the internet.

Speech synthesis it would appear has lost out to the success of speech coding. It does not seem to me that there will be the same range of applications as for recognition. The widespread display of data on low-cost displays will obviate the need for synthesis in many applications. However there are several areas where synthesis may be useful when a display is not available to read text or it is dangerous to do so.

If *speaker verification* is to have applications outside the forensic area it must be carried out in an unobtrusive way as an integrated part of the system. Speaker verification will be replaced by the idea of speaker validation where a speaker's identity is established by using a text independent speaker verifier using all of the speech spoken in a transaction.

6 CONCLUSIONS

Speech technology now has a number of outstanding successes to its credit particularly the widespread use of speech coders and speech to text on PCs. It has a very bright future but this may be dimmed and delayed if workers in the field do not apprehend the rapid changes that are occurring in the rest of information technology. These changes will result in more sophisticated user interfaces which should contain speech technology. They will only do so if speech technologists are aware of other developments in information technology and direct their efforts towards designing applications which combines well with the rest of the system.

7. REFERENCES

Because of the speed at which the technology is developing I have chosen to cite the references in the form of Internet addresses so that the reader can access the most up to date information.

[1]http://www.itu.int
[2]http://www.dragon-systems.com
[3]http://infopad.eecs.berkeley.edu
[4]http://www.alcatel.com
[5]http://www.lsilogic.com
[6]http://www2.echo.lu/langeng/test/en/
[7]http://palmpilot.3com.com/
[8]http://www.bluetooth.com

ISBN 1 901656 14 4

ISSN 0309 - 8117