Spoken Language and e-inclusion.

Prof Alan F Newell, MBE, FRSE.
Applied Computing, University of Dundee, Dundee, Scotland
afn@computing.dundee.ac.uk.

Abstract

Speech technology can help people with disabilities. Blind and non-speaking people were amongst the first to be provided with commercially available speech synthesis systems, and, to this day, represent a much higher percentage of users of this technology than their numbers would predict.

Speech synthesis technology has, for example, transformed the lives of many blind people, but the success of speech output to allow blind people to word processes, browse the web, and use domestic appliances should not to lull us into a false sense of security. In the main, these users were young, aware of their limitations, and of the substantial potential impact of such technology on their life styles, and were generally highly motivated to make a success of their use of the technology.

The speech community needs to be aware of the major differences between the young disabled people who have found speech technology so useful, and the other groups of people who are excluded from "e-society". An example is older people. These have a much greater range of characteristics than younger people and these characteristics change more rapidly with time. Very importantly for speech technologists, most older people possess multiple minor disabilities, which can seriously interact, particularly in the context of a human machine communication. In addition, a relatively high proportion of older people also have a major disability.

1. Introduction

The motivation behind the workshop for which this paper has been prepared is the premise that Spoken Language Processing has the potential to address the issues of 'e-inclusion' of users groups such as those with special needs including disabilities and illiteracy, children and older people and the technologically disadvantaged (in either the developed or developing world).

We already have some excellent existence proofs that speech technology can help people with disabilities. Blind and non-speaking people were amongst the first to be provided with commercially available speech synthesis systems, and, to this day, represent a much higher percentage of users of this technology than their numbers would predict. Automatic Speech Recognition (ASR) technology also proved popular amongst some physically disabled people - although to a lesser extent than speech synthesis technology. In fact, this author moved into the field of developing systems for disabled people, in the late nineteen sixties, when he was researching into Speech Recognition technology [1].

Speech synthesis technology has, for example, transformed the lives of many blind people, but the success of speech output to allow blind people to word processes, browse the web, and use domestic appliances should not to lull us into a false sense of security. In the main, these users were young, aware of their limitations, and of the substantial potential impact of such technology on their life styles, and were generally highly motivated to make a success of their use of the technology. Speech technology proved very successful for these users particularly because they were prepared to put significant effort into understanding what the system said. In this sense, blind people were ideal users for this emerging technology.

2. E-inclusion and older people

The above examples, however, were a special case which is unlikely to be repeated for other e-inclusion issues. For example, one group of people, who suffer from e-exclusion, are older people. Their reduced functionality, often combined with a distrust of technology, apparently make them an obvious target for communication with machines based on speech technology. In addition the demographic trends throughout the world show that this group of people is of growing importance. The UK Office of National Statistics quote the following figures and predictions:

In 1995 there were 2.3 million people aged over 80
by 2020 there will be 3 million
by 2030 there will be 4 million
by 2040 there will be 4.8 million
by 2050 there will be 5.5 million people aged over 80

In addition to their increasing share of the population throughout the world, this group represent a portion of the population of the developed world which contains large numbers of people with significant wealth, disposable income, and leisure time, and thus offer potentially large markets for technology [2].

The speech community, however, needs to be aware of the major differences between older people and the

young disabled people who have found speech technology so useful. Firstly older people have a much greater range of characteristics than younger people and these characteristics change more rapidly with time. In addition, and very importantly for speech technologists, as well as a relatively high proportion of older people having a major disability, most older people possess multiple minor disabilities, which can seriously interact, particularly in the context of a human machine communication [3].

3. Speech Synthesis systems for Older People

Spoken output is an obvious, and well tried, solution to problems of a blind person reading text on a screen. In addition to the presented visual impairment, however, an older user will most likely also have a hearing impairment, which will make their ability to understand synthesised speech more difficult – particularly in noisy environments. Even if the speech can be presented at a loud enough level to be heard - and this is may not always possible - understanding this speech will provide a greater cognitive load than would be exerted on a person with normal hearing. This is further complicated, by another effect of old age - the slowing down of cognitive processing. Older users may be able to hear the speech (sometimes with some difficulty), but may not have enough spare cognitive capacity to adequately decode the messages contained in it. Newell, Carmichael et al [4] discuss the cognitive functioning of older people and how this relates to the design of interfaces to information technology.

A further indication of reduced capacity in older people is the effect which is noticed when different qualities of speech are introduced into a message - for example telephone directory enquires, where the telephone number is synthesised but the pre-amble is recorded speech. It has been found that older peoples' ability to listen to such mixed quality speech is less than that of younger people. This is probably because of the cognitive load involved in "resetting one's recognition algorithms" to cope with a changing speech qualities.

Older people with good vision use lip reading to assist with poor hearing (sometimes leading to the slightly odd statement "I cannot hear you, let me put on my glasses"), and for sighted people an avatar lip synchronised to the speech can improve intelligibility. In the development of any service based on such a provision, however, it must be remembered that the accuracy of movement and time synchronisation of the lips and the avatar are crucial. Minor temporal misalignments, or rendition accuracy will increase cognitive load on the user, and, if the quality drops below a certain level, the avatar's lips may impede understanding of the speech rather than assist it. An example is the so-called "McGurk effect", where for example if /duh/ is presented on the lips and /puh/ presented auditorally, the combined effect has been found to be a perception of /buh/.

A further cognitive loading is caused by data which has been originally designed to be presented visually. Understanding such data in a spoken form is a learned activity and requires additional cognitive effort and a good memory (even in an up-market restaurant where there is no printed menu!). When synthesised speech is used, and data which is more complex than a simple short list, the task can be very difficult, particularly for an older person who is un-used to this activity. Additional problems for speech output include the most appropriate way of speaking out tabular data, and how to annotate pictures most effectively. In addition, person to person spoken dialogue is usually a two way process, with very subtle rules about turn taking, how to indicate you wish to interrupt, and how to actually interrupt. There should be no need to worry about being polite to a computer, but some users may be inhibited because of social custom. Care needs to be taken in the management of such spoken dialogue, even if the response is via a keyboard. Carmichael [5] discusses these and similar issues in relation to the challenges of voice control of an electronic programme guide for television.

4. Speech recognition systems

The ideal solution, on the face of it, for an older person is two-way spoken dialogue, but speech recognition is a very much more technically complex issue than speech synthesis, and despite a great deal of research it has yet to reach the potential many people believe it to have. Proponents of speech understanding systems throughout the years of its development, have had a tendency to underestimate the challenges of a particular application and have articulated (often grossly) over-optimistic views of the performance of the technology and its ease of use. The author commented on this some years ago in his paper " Speech the natural modality for man-machine interaction" [6]. Speech processing technology has improved enormously since that paper was written, but the basis premise remains that there is much more to designing a speech system than recognising what people say, and the performance of even current systems are less than adequate for many application areas.

For similar reasons to the blind users of speech synthesis, some severely physically disabled people have been more prepared than mainstream users to battle with the training schedules of speech recognition systems. ARS machines have even been found to work

satisfactorily with people with articulatory dysfunction, provided they are able to make differentiable sounds. Again, however, these disabled people were a highly motivated group of mainly younger people. Older users provide similar challenges to those mentioned previously in this paper of multiple minor disabilities and cognitive overload. Other challenges for ASR system include very low volume and/or poorly articulated speech of some frail older people, and their being less used to speaking into microphones than younger people.

Many older people are less patient with machines which go wrong, than younger people, and are inhibited from trying alternative approaches when they are not sure what the problem is. They will be less able to cope with a speech recognition machine which makes errors, particularly when there is no obvious connection between what one said and what the machine recognised. Older people are also more likely to become confused by erroneous responses to their commands, and less likely to remember particular, and idiosyncratic dialogues structures which a system may require to work accurately.

The relative technical naivety of many older people can exacerbate problems which have been seen to appear in mainstream areas. A recognition system which appears to be working well, can have a deleterious effect on the users performance by causing them to relax into normal spoken behaviour and conventions. Underling this activity is often an unconscious assumption that the system has significantly more intelligence and the other characteristics of a human being than it in fact does. This can lead to a reduction in the precision of the spoken language which is used. This can provide significant challenges for systems, particularly those which depend on a closely controlled dialogue structure. In experiments concerning electronic programme guides, for example, Carmichael [7] reported that, even when the dialogue was severely constrained to answering the simple question "now" or "later", young people would respond appropriately, particularly if the appropriate words were emphasised in the question. Older people, on the other hand, were inclined to reply with a polite sentence, "Oho I would like to watch something now please". {Expanding the dialogue structure to cope with this could lead to absurdities such as "please say now now, or later now"}. It is not easy to use spoken dialogue to indicate very precisely how a spoken response should be formulated. This is not normal behaviour for human beings, and thus very difficult to ensure with untrained users. Turn taking behaviour has been mentioned in relation to speech output devices; when the speech dialogue with the machine is two way, an increased tendency to fall into natural turn-taking behaviour, rather than the more artificial behaviour of the machine, could lead to problems. In addition, a great deal of human machine activity involves pointing – particularly since the advent of the WIMP interface, but pointing with one's voice is not a natural activity. Trying to replace mouse activity with a spoken dialogue presents major challenges to a system designer.

Most of these challenges are not substantially different from those presented by younger users of spoken language technology, but older people often present these challenges in a more extreme form – in this sense they are a good user population to consider, because they can provide more of a "test to destruction" of the technology than often occurs prior to release of a new system. Everyone suffers cognitive overload to some extent, and can fail to use a piece of technology because of this effect. An attempt to reduce this effect by developing systems for people for whom cognitive overload is more regular occurrence can lead to better technological solutions for everyone.

5. Other "extreme" users

Although the above points have been focussed on older people, the challenges presented by those in technologically disadvantaged countries, and the users of minority languages are not dissimilar. This is particularly true in terms of the amount of extra cognitive activity which needs to be applied by these groups of people when using speech systems for both input and output of data, and the fact that that they will have less spare cognitive capacity than a person who is using their own language within a technologically mature society.

6. Value for mainstream Spoken Language Research

A focus on the needs of older or disabled people can lead to data and systems which are more widely useful [8]. The most ubiquitous example of a "speech system" in this context is the cassette tape recorder. This technology was originally developed by a company producing talking books for the blind – and at the time the engineers claimed this technology would never be popular because of the poor sound quality. A more appropriate example for the current speech research community was the initial research into emotion within speech synthesis [9]. This was motivated by the needs of non-speaking users of speech synthesis systems (Augmentative and Alternative Communication (A.A.C.) Systems who need to express emotions. This research pre-dated, and formed the foundation of, a more general examination of the need to introduce emotion into speech synthesis systems. The commercial trend of moving from formant-based synthesisers to concatenative synthesizers, based on recorded speech

however has introduced new challenges. Concatenative systems provide improved speech quality and naturalness, which can be very valuable for many applications, but they are less flexible than formant synthesisers. There is far less scope for adding the speech variations needed to convey speaking styles and emotions or for offering a synthesiser with multiple voice personalities. This presents a particular problem for AAC users who wish to be able to chose the personality and characteristics of the voice they use (early AAC systems used US male voices, the only ones then available, which was less than satisfactory for a non-US female AAC user !). Many AAC systems having identical speech quality also cause confusion when more than one user is present in a conversation. Emotional tone in synthetic speech could also be useful for adding extra information to speech signals to assist blind or illiterate people in reading text. Basic attempts to add emotion to concatenative synthesis have had some success, but there are serious limits to how far the concatenative systems can give appropriate speech variation [10]. Newell et al [11] discuss other ways in which natural language processing research can be applied to the problem of developing speech output systems for non speaking people.

7. Conclusions

Speech Technology has had, and will continue to have, a major impact on the lives of excluded people, but only if it designed in a holistic manner within an inter-disciplinary research culture [12].

The major lesson which can be learned from developments of systems designed with the older user in mind is the importance of being fully informed about the actual characteristics of the user, a full understanding of the task which they wish to perform and a clear understanding of the real environment (acoustic, social, visual) in which such task will be performed [13]. There are innumerable examples where this philosophy has not been followed and the end product has been less than successful.

When speech systems are designed and tested by and with young technically aware people, in a laboratory setting, they appear to work very well, but are not always a real commercial success. The use of older and disabled people as role models of users, and also within the testing and evaluation of systems, can provide the spur for engineers to develop systems which really work in real environments with real people.

This design philosophy will benefit older and disabled people, but also, a consideration of extra-ordinary human beings, can lead to better systems for all users and also important research data and insights.

8. References

1. A.F. Newell, Man machine communication using spoken morse code, *Intl. J. of Man-Machine Studies*, **2**, 351-362, 1970.
2. F. Newell, "Assistive technology - the older people's perspective", in: *Proceedings of 3rd TIDE Congress, Keynote Speech - Improving the Quality of Life for the European Citizen (4 June 1998)* **4**, pp.xlviii-lii.
3. A. F. Newell and P. Gregor (2002), "Design for older and disabled people - where do we go from here?", *Universal Access in the Information Society* 2(1) (2002) pp.3-7.
4. A. F. Newell, A. Carmichael, P. Gregor and N. Alm, "Information technology for cognitive support", *The Human-Computer Interaction Handbook* **2** (2002) pp.464-481
5. A.R. Carmichael, “Talking to your TV. *UsableiTV,* issue 3 pp 17-20, 2003 (see www.itv.network.org)
6. A F Newell. *Proc. 1st IFIP Conf. On Human-Computer Interaction – Interact’84,* London, 231-238, Sept. 1984, Ed. B Shackel, Pub. North Holland, 1985)
7. A.R. Carmichael, Applied Computing, University of Dundee, Private Communication 2003.
8. A F Newell and P Gregor, Extra-ordinary human-machine interaction – what can be learned from people with disabilities?,. *Cognition Technology and Work,* **1(2),** 78-85, 1999.
9. I.R. Murray and J.L. Arnott, "Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion", *Journal of the Acoustical Society of America,* 93(2), February 1993, pp. 1097-1108.
10. I.R. Murray, "Rule-based Emotion Synthesis using Concatenated Speech", *Proceedings of the ISCA workshop on Speech and Emotion,* Newcastle, Northern Ireland, 5-7 September 2000, pp. 173-177
11. A. F. Newell, S. Langer and M. Hickey, "The role of natural language processing in alternative and augmentative communication", *Natural Language Engineering* **4(1)** (1998) pp.1-16.
12. A. F. Newell and P. Gregor, "User sensitive inclusive design in search of a new paradigm", in: *CUU 2000 First ACM Conference on Universal Usability (USA 2000)* (ed. J. Scholtz and J. Thomas) pp.39-44.
13. A. F. Newell and P. Gregor, "Human computer interfaces for people with disabilities", *Handbook of Human Computer Interaction* (1997) pp.813-824.