Strategic Promotion of Ageing Research Capacity

What Makes Synthetic Speech Difficult for Older People to Understand?

Maria Klara Wolters,
Pauline Campbell,
Christine DePlacido,
Amy Liddell & David Owens

Meeting the challenges of an ageing society

Funded by

EPSRC Engineering and Physical Sciences Research Council

bbsrc biotechnology and biological sciences research council

October 2008

What Makes Synthetic Speech Difficult for Older People to Understand?

Dr Maria Klara Wolters, Pauline Campbell, Christine DePlacido, Amy Liddell & David Owens
University of Edinburgh

Computer-generated voices are used and encountered more and more frequently in everyday life. They can be found not only in automated call centres, but also in satellite navigation systems and home appliances. However, older people can find computer-generated speech difficult to understand, and some simple changes to the voices used could make an immense difference to the lives of many older people. This would make services which use computer-generated voices more accessible to all. A study was carried out to investigate the effect of auditory ageing on the ability to understand both computer-generated (known as synthetic) and natural voices. Major differences in the ability to understand synthetic speech were found between younger and older people. These differences were shown to be in relation to the type of voice, the language used, and the complexity and unfamiliarity of key words. From these findings, guidelines for improving synthetic speech have been produced.

Key Findings

- Natural auditory ageing was found to affect participants' hearing performance even when it was at levels that are normally regarded as healthy.
- Both older and younger participants had no problems understanding synthetic speech about things they were familiar with, as long as it used well-known words and phrases. However, they struggled with unfamiliar, long, and complex words.
- Whether participants were able to remember speech correctly did not depend so much on their memory, but on their ability to hear frequencies between 1 and 3 kHz. These frequencies are in the lower part of the middle range of frequencies which the ear can hear. They contain a large amount of information about the identity of speech sounds.
- Central auditory processing (processing of sounds by the brain) appeared to play a more important role in the ability to understand natural speech, whilst peripheral auditory processing (processing of sounds in the ear) appeared to be more important for synthetic speech.
- The findings from the study have enabled some initial guidelines to be produced for adapting unit selection synthetic speech (a technique commonly used by modern speech synthesis systems) for older users with mild to moderate hearing loss.

Introduction

The Issues

Many older people have difficulty making telephone calls to call centres or following instructions from the voices used in in-car navigation systems. Part of the problem may be the computer-generated voices which are often used in these applications. Older people can find these voices hard to understand, which can be very frustrating and cause anxiety. This research project was concerned with making computer-generated voices easier for older people to understand and more pleasant to listen to.

The Background

Modern speech synthesis systems are based on **unit selection technology**; that is they generate speech by searching for the appropriate units (words, syllables, sequences of sounds) from a large pool of data, and then join these units together. Although there have been considerable advances in this technology in recent years, even the leading systems can be difficult to understand. For example, when the units selected do not join together well, this forms clicks; and if units are shorter or longer than they need to be, the speech rhythm becomes distorted. This can be a particular problem for older listeners because, as a group, they tend to have a range of hearing difficulties due to age-related anatomical and physiological changes to the ear.

These changes can lead to many problems including hearing loss in the higher frequencies, and difficulty understanding speech against a background of noise. Some older adults also have difficulty following distorted speech.

The Aims

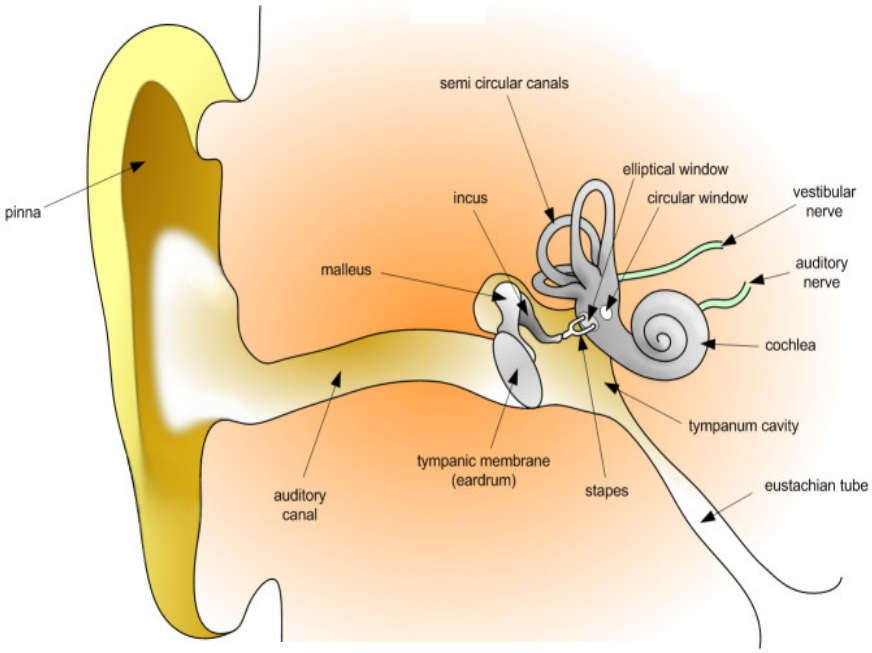
There were three main aims of the study:

- to investigate the effect of auditory ageing on the ability to understand synthetic speech;
- to assess older peoples' attitudes to synthetic speech;
- to design a programme of research into adapting synthetic speech to make it easier for older people to understand.

The Study

The study consisted of three parts:

- tests to investigate both auditory ageing and working memory aspects of cognitive ageing;
- a short experiment testing the ability of people of all ages to understand unit selection speech synthesis;
- a short, structured interview in which participants were asked to assess a synthetic voice, and to give their views on using synthetic speech in home-care applications.



The Ear

http://en.wikipedia.org/wiki/Image:HumanEar.jpg

Data collection

Detailed measurements of hearing loss due to problems with the outer, middle, and inner ear (peripheral hearing loss) were made. These measurements are important because processing of sound in the ear (**peripheral auditory processing**) affects further processing by the brain (**central auditory processing**).

The interpretation of sounds (**auditory stimuli**) may also be affected by general cognitive ageing. In general, our ability to process new information is reduced as we age. Deciphering difficult auditory stimuli takes up resources which could otherwise be used for processing or storing the information that these stimuli contain. Hence, problems with understanding speech may not be solely due to auditory problems, but could be associated with cognitive decline.

To control for potential interactions between auditory and cognitive ageing, the participants' working memory span was assessed. Working memory was chosen because it is used to briefly store the information needed for cognitive processing. Therefore, it plays an important role in understanding and recalling information.

The participants

Forty four participants were recruited from the local community, from the Queen Margaret University, and from the University of Edinburgh. They were split into three age groups: 60-70 years old, 50–60 years old and "younger people" (20-30 years). Although the 50–60yr old group had significantly less clinical hearing loss compared to the 60-70yr old group, they were found to have clear signs of auditory ageing.

Age Group	**Older** 60–70	50–60	**Younger** 20-30
Number in group	12	20	12
% female	66%	40%	84%
% male	34%	60%	16%
% with hearing problems in noisy environments	75%	45%	8%

The Participants

Assessment tests

A range of tests were used to measure the hearing of the participants.

- **Traditional pure tone audiometry** (PTA) is a measurement of hearing threshold where participants respond to a **pure tone stimulus** (a single frequency sound). It is used to assess hearing loss and uses pure tones up to 8MHz.
- **Extended high frequency audiometry** (EHF) covers the whole range of higher frequencies from 9–20 kHz. It includes frequencies that can typically only be heard by healthy young people. Combining EHF and PTA gives a clear picture of cochlear sensitivity across the whole frequency range.
- **Otoacoustic emissions** (OAEs) are sounds produced by the ear. They can either be produced spontaneously, or by using carefully selected sounds, such as clicks or two pure tones that are presented at the same time. OAEs can predict type and severity of hearing loss.

The full assessment comprised:

- a detailed questionnaire covering subjective assessment of hearing loss and relevant aspects of participants' medical history;
- pure tone audiometry (PTA) in quiet conditions, separately for each ear at frequencies from 0.25 to 8 kHz;
- extended high frequency audiometry (EHF) in quiet conditions, separately for each ear at frequencies from 9 to 20 kHz;
- speech audiometry using word lists presented in quiet conditions;
- otoacoustic emissions (OAEs);
- Random Gap Detection Test (RGDT), which detects the point at which two stimuli become recognised as one rather than two;
- Working Memory Span test, which assesses how much information can be retained in the short term memory.

The speech synthesis experiment

The speech synthesis experiment took advantage of a unique collaboration with Cereproc, a spin-off company from the Centre for Speech Technology Research in Edinburgh. Cereproc has developed a state-of-the-art, commercial, unit selection speech synthesis system which offers an excellent Scottish-English voice, known as “Heather”.

Thanks to Cereproc’s co-operation, the speaker who had provided the speech data for “Heather” agreed to record all the experimental stimuli. This provided an unparalleled match between the synthetic and the human voice, since the only difference between the two was the additional processing which was required to generate the synthetic speech. Almost all previous research has used human voices that are likely to have been quite different in both voice quality and tone to the synthetic voice to which they were compared.

The tests

Participants heard a total of 32 different reminders: 16 meeting reminders specifying a time and a person; and 16 medication reminders specifying a time and a medication name.

In half of the reminders, the time came first in the sentence. In the other half, the time was presented second. The medication names were designed by recombining parts of existing medication names, so that all names were guaranteed to be unfamiliar. The person names were designed to be easily confused. They typically consisted of a consonant-vowel-stop sequence and were chosen to ensure the existence of similar names with which they could be confused, for example “Rob”, “Bob” and “Ron”.

Reminder Type	Pattern	Synthetic	Human	Total
Medication	At T, you need to take your M	4	4	8
Medication	You need to take your M at T	4	4	8
Meeting	At T, you’re meeting P	4	4	8
Meeting	You’re meeting P at T	4	4	8

Sentence Patterns for Speech Reminders

All stimuli (sounds) were presented via headsets with the loudness adjusted to the individual’s degree of hearing loss. After hearing a reminder once, participants were asked a related question such as “When do you need to take your medication?” or “Who are you meeting?”.

Results

General results

Three key factors were found to have a major impact on the participants’ ability to remember the words they had heard correctly: stimulus category (person, time, or medication); voice (synthetic or human); and position of the required information in the reminder (first or second).

- Participants found person names and times far easier to remember than medication names. All person names were short, relatively familiar, and their phonological structure was simple. The medication names, on the other hand, were specifically designed to be unfamiliar, long, and phonologically complex.
- Only the complex medication names were more difficult to remember when given by synthetic speech. Participants remembered times and person names well no matter what the voice.
- Items in second place within the sentence were found to be easier to remember than items in first place.

Contrary to expectations, working memory span was not found to be associated with participants’ ability to understand the synthetic speech. Instead, working memory span was associated with the participants’ ability to understand the human voice.

- The impact of auditory ageing

The study examined whether synthetic speech was more easily understood by those with better hearing. No associations were found between the extent to which speech was understood and the traditional hearing screening threshold frequencies of 0.5, 1, 2, and 4 kHz. As a consequence, experimental designs which measure only the four traditional screening frequencies may miss crucial information that might explain some of the variability in understanding.

However, it is important to note that effects of auditory ageing were found even in participants with clinically normal hearing.

The study also examined whether otoacoustic emissions (OAEs) could predict whether synthetic speech could be understood. OAEs are thought to be linked to the outer hair cells of the cochlea, which are involved in increasing frequency selectivity and sensitivity, acting as a kind of amplifier in the ear.

If damage to these affects how well participants can understand synthetic speech, then a filter to amplify the frequency could easily be used to solve the problem. However, OAE levels did not correspond to participants' level of understanding of synthetic stimuli.

Importantly, the study found that auditory ageing affects the understanding of natural and synthetic speech differently. Central auditory processing (undertaken by the brain) was found to be more important in the participants' ability to understand natural speech than their ability to understand synthetic speech. Therefore any problems with this part of the hearing process are likely to affect peoples' understanding of normal speech. Peripheral auditory processing (undertaken by the ear) was found to be most important in the participants' ability to understand synthetic speech. Any problems with this part of the hearing process are likely to affect peoples' understanding of synthetic speech.

Audiological findings

Seven of the 18 older participants with no clinical hearing loss in either ear reported difficulties hearing in noisy environments. These older adults also showed significantly elevated pure-tone thresholds. Despite these differences, these adults would be classified as having perfectly healthy hearing by standard UK testing procedures. Funding has been secured to explore this issue further.

Further analysis of the audiological data suggested that participants' ability to remember medication names correctly might depend more on their ability to hear frequencies between 1 and 3 kHz than on their memory. These frequencies contain a large amount of information about the identity of speech sounds.

The study also assessed whether the Random Gap Detection Test (RGDT) was a useful tool for distinguishing between older and younger adults' ability to detect gaps between two sounds (auditory stimuli). A significant difference between the two groups was found for the lowest tonal stimuli (0.5 kHz), with only the younger group showing the ability to detect a gap between two pure tones at this frequency.

The interviews

Overall, participants found the task of listening to the voices fairly easy. In the interviews they noted that clarity of diction was more important than accent. Any accent present in the voice should be fairly soft to ensure that the voice can be understood easily. Some noted that the computer-generated version of the voice seemed to be a bit unclear or to swallow the odd syllable. Surprisingly, quite a few of the participants found it difficult to tell the difference between the human and the computer-generated version of the voice.

Guidelines

Guidelines

The following guidelines for the use of synthetic speech were produced from the findings of this study:

- It is more important to speak clearly than to speak slowly. The frequency range that was particularly strongly associated with the ability to understand often contains information that is expressed in short changes from one sound to another. These changes (transitions) should not be shortened or distorted unnecessarily.
- Use phrasing to make information more noticeable and obvious. In some particularly difficult medication names, the initial and final sounds were too short or distorted during the synthetic speech, because of problematic transitions between these sounds and the surrounding material. Placing pauses between such difficult information and the surrounding words cannot only eliminate these troublesome transitions, but also highlight that the information contained in the phrase is important.
- Use familiar words and phrases. The only times when clear differences in understanding emerged between natural and synthetic speech was when complex, unfamiliar words and complicated sentences were used. It is good design practice to avoid this type of language when designing system prompts.

In the spirit of inclusive design, these guidelines represent recommendations for best practice that will benefit a large number of users ranging from those who have no problems, to users with moderate hearing loss. Quite explicitly these do not support “dumbing-down” the content of messages which can be counterproductive, but merely the computer equivalent of “clear diction”. Future work will validate and test these guidelines.

Impact

The findings have already attracted attention from external groups. British Telecom has expressed an interest in the guidelines, and a research paper about the study has been adopted as a reference document for the American National Standards working group on text-to-speech technology.

The audiological work has led to a follow-on project supported by a grant from the Queen Margaret University Centre for the Promotion of the Older Person’s Agenda, *“The Use and Abuse of Auditory Profiles”* (lead investigator: Christine DePlacido).

The tests used for this study, which include the brief and highly portable speech synthesis experiment, are now being used in other studies currently underway at Queen Margaret University’s Section of Speech and Hearing Sciences. The group of tests described will also be used for the Auditory Profiles project. Thus, this study has enabled the accumulation of data and the creation of a potentially large database on auditory function of older people.

The Research Team

Dr Maria Wolters, Research Fellow
Centre for Speech Technology Research
School of Informatics, University of Edinburgh, Informatics Forum,
10 Crichton Street, Edinburgh EH8 9AB
mwolters@inf.ed.ac.uk

Pauline Campbell, Lecturer
Speech and Hearing Sciences
Queen Margaret University
Musselburgh, EH21 6UU

Christine DePlacido, Audiologist
Speech and Hearing Sciences

Amy Liddell, Audiologist
Centre for Speech Technology Research
School of Informatics

David Owens, Audiologist
Centre for Speech Technology Research
School of Informatics

The Study

The study received financial support from SPARC of £17,541 and ran for 12 months ending in November 2007. Additional support was provided by the University of Edinburgh.

More information about the study can be found on the SPARC website www.sparc.ac.uk and obtained directly from the investigators.

Bibliography and Key References

Hunt A., Black A. W. 1996. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. *ICASSP-96.* Atlanta, Georgia. 1:373–376.

Keates S., Clarkson J. 2004. *Inclusive Design.* London: Springer.

Silman S., Silverman C. 1991. *Auditory Diagnosis: Principles and Applications.* Thomson Publishing.

Wolters M., Campbell P., DePlacido C., Liddell A., Owens D. 2007. Making Synthetic Speech Accessible to Older People. *Proceedings of the Sixth ISCA Workshop on Speech Synthesis,* Bonn, Germany.

Acknowledgements

The research team is indebted to the individuals who took part in the study, and to Matthew Aylett and Christopher Pidcock at Cereproc.

Disclaimer

SPARC

SPARC is a unique initiative supported by EPSRC and BBSRC to encourage the greater involvement of researchers in the many issues faced by an ageing population and encountered by older people in their daily lives. SPARC is directed, managed and informed by the broader community of researchers, practitioners, policy makers and older people for the ultimate benefit of older people, their carers and those who provide services to older people.

SPARC pursues three main activities:
Workshops to bring together all stakeholders interested in improving the quality of life and independence of older people.
Advocacy of the challenges faced by older people and an ageing population and of the contribution of research to improving quality of life. SPARC is inclusive and warmly welcomes the involvement of everyone with a relevant interest.
Small Awards to newcomers to ageing research, across all areas of design, engineering and biology and at the interfaces relevant to an ageing population and older people. In 2005 and 2006 SPARC received 185 applications for support in response to two invitations for competitive proposals of which 34 were supported.

Executive Summaries

SPARC is supporting its award holders through funding, mentoring, a prestigious dissemination platform, professional editorial assistance, international activities and provision of contacts. Each of the projects has been small, yet the enthusiasm for discovery, and impatience to contribute to better quality of life for older people, has more than compensated for the very limited funding which was provided.

This executive summary is one of a series highlighting the main findings from a SPARC project. It is designed to stand-alone, although taken with summaries of other projects it contributes to a formidable combination of new knowledge and commitment by newcomers to ageing research, with a view to improve the lives of older people. This is a tangible contribution towards ensuring that older people receive full benefit from the best that research, science and technology can offer.