Making Speech Synthesis Accessible for Older People

Maria Wolters\textsuperscript{1}  Pauline Campbell\textsuperscript{2}  Christine DePlacido\textsuperscript{2}
Amy Liddell\textsuperscript{2}  David Owens\textsuperscript{2}

\textsuperscript{1}Centre for Speech Technology Research
University of Edinburgh

\textsuperscript{2}Speech and Hearing Sciences
Queen Margaret University

6\textsuperscript{th} ISCA Speech Synthesis Workshop, Bonn, Germany
Overview

1. Making TTS Accessible to Older People
2. Pilot Study: Reminders in the Home
   - Experiment Design
   - Participants and Assessments
3. Results
   - Effect of Stimuli
   - Effect of Cognitive Ageing
   - Effect of Auditory Ageing
4. Conclusion
   - Summary
   - Future Work
Overview

1. Making TTS Accessible to Older People
2. Pilot Study: Reminders in the Home
   - Experiment Design
   - Participants and Assessments
3. Results
   - Effect of Stimuli
   - Effect of Cognitive Ageing
   - Effect of Auditory Ageing
4. Conclusion
   - Summary
   - Future Work
Overview

1. Making TTS Accessible to Older People
2. Pilot Study: Reminders in the Home
   - Experiment Design
   - Participants and Assessments
3. Results
   - Effect of Stimuli
   - Effect of Cognitive Ageing
   - Effect of Auditory Ageing
4. Conclusion
   - Summary
   - Future Work

Wolters, Campbell, DePlacido, Liddell, Owens
Speech Synthesis for Older People
Overview

1. Making TTS Accessible to Older People
2. Pilot Study: Reminders in the Home
   - Experiment Design
   - Participants and Assessments
3. Results
   - Effect of Stimuli
   - Effect of Cognitive Ageing
   - Effect of Auditory Ageing
4. Conclusion
   - Summary
   - Future Work
What does Accessibility Mean?

- Older people can **understand** the voice easily.
- Older people **like** the voice:
  - its dialect (Zajicek)
  - its persona (Nass)
  - the emotion it projects (Wolters/Boye/Oberlander, just completed)

This study focusses on **intelligibility**.
Are State-of-the-Art Voices Easy Enough to Understand?

What is state-of-the-art in the Human Factors, Rehabilitation and Augmentative and Alternative Communication literature?

- In 2007, Roring, Hines, and Charness claimed in “Human Factors” that “Synthetic speech fidelity must be improved significantly before becoming truly useful for the older population”.

Which system did they test?

- **Black and Taylor 1997, American English Diphone voice.**
In a 2006 study published in “Augmentative and Alternative Communication”, Alampasutra, Kohnert, Munson, and Reichle decided to use a system

“[...] because it is still one of the most frequently used speech synthesis packages used with electronic communication aids and educational software.”

Which system did they use?

DECTalk.
What Do We Know About Older People’s Problems With Understanding TTS?

- **a lot** about problems with formant synthesis:
  - lack of detailed acoustic information (Duffy and Pisoni 1992)
  - speed too high (Sutton et al. 1995)
  - poor prosody (Paris et al. 2000)

- **a little** about problems with diphone synthesis
  - slowing it down doesn’t work as well as slowing down DECTalk (Roring et al. 2007)
  - human voices are far more natural (Lines and Hone 2006)

- **nearly nothing** about problems with unit selection synthesis
  - People with hearing problems have more trouble recognising words (Langner and Black 2005)

- **nothing** about problems with HTS.
Common Problems with Literature

- DECTalk is speech synthesis (dominant market position)
- Speech synthesis is monolithic: Results for formant synthesis can be extrapolated to concatenative synthesis

Key Gap: Very little research to ensure that state-of-the-art technology is inclusive
Research Questions for Pilot Study

- What makes an utterance generated by unit-selection TTS difficult to understand for older people?
- How can we adapt our systems to support older people?

Systematically investigated in this study: What are the relevant aspects of auditory ageing?
Making TTS Accessible to Older People

Pilot Study: Reminders in the Home
- Experiment Design
- Participants and Assessments

Results
- Effect of Stimuli
- Effect of Cognitive Ageing
- Effect of Auditory Ageing

Conclusion
- Summary
- Future Work
Task: Understanding Reminders

- Participants first hear a reminder, followed by a question about the reminder.
- Participants have 10 seconds to answer verbally.
- 8 training items before main experiment, 4 per voice.
- Training Materials (Sound Example)
Material

2 Types of Reminders:

- 16 meeting reminders
  (You are meeting P at T. or At T, you are meeting P.)
  Person names CVC monosyllables, easily confundable

- 16 medication reminders
  (You need to take your M at T. or At T, you need to take your M.)
  Medication names constructed from morphemes of real names, 3-4 syllables long

Questions:

- Which medication?
- Which person?
- At what time?
Stimuli

- **Reminders:** human or synthetic voice
- **Questions:** human voice
- **Two voices (Sound Example):**
  - synthetic voice: Cerevoice “Heather”
  - human voice: “Heather”
- Exceptionally close match between human and synthetic voice: same post-processing, same speaker
- Stimuli grouped into 4 balanced lists. Each listener heard one list
Scoring

Responses were scored as **correct** if they were an acceptable pronunciation of the target, modulo dialect. **Error categories:**

- **phoneme errors:** insertion, deletion, or replacement of a single phoneme
- **syllable errors:** $>1$ phoneme errors in a syllable; syllable insertion/deletion
- **word errors:** word omitted or inserted
1. Making TTS Accessible to Older People

2. Pilot Study: Reminders in the Home
   - Experiment Design
   - Participants and Assessments

3. Results
   - Effect of Stimuli
   - Effect of Cognitive Ageing
   - Effect of Auditory Ageing

4. Conclusion
   - Summary
   - Future Work
Participants

- 12 younger, 32 50+ (20 50–60, 12 60–70)
- younger people screened to exclude people with low-frequency hearing loss
- older people screened to exclude people with hearing aids and people with severe losses
Cognitive Assessments

- Working Memory Span (WMS; Unsworth et al. 2005)
  - relevant for task
  - good proxy of cognitive ageing
- Prospective and Retrospective Memory Questionnaire (Crawford et al.)
Auditory Assessments

Tests cover both peripheral and central auditory processing.

**Pure-Tone Audiometry:** 0.25, 0.5, 1, 2, 3, 4, 6, 8 kHz

**Extended High Frequency Audiometry:** 9, 10, 11.2, 12.5, 14, 16, 18, 20 kHz

**Speech Audiometry:** set of normed CVC word lists

**Gap Detection:** size of smallest gap between two clicks that can be detected reliably (Random Gap Detection Test, Keith 2000)
Outline

1. Making TTS Accessible to Older People
2. Pilot Study: Reminders in the Home
   - Experiment Design
   - Participants and Assessments
3. Results
   - Effect of Stimuli
   - Effect of Cognitive Ageing
   - Effect of Auditory Ageing
4. Conclusion
   - Summary
   - Future Work
What Makes Information Easy to Remember?

General Results

Score $\leftarrow$ Position $\times$ Category $\times$ Voice

- Information is time (when?) or person (what?) ($p<0.000001$)
  - times and person names are short and familiar
  - medication names complex and unfamiliar

- Information presented in human voice ($p<0.0001$)
  - Interaction with category: worst for medications

- Information presented in second position ($p<0.05$)
What Makes a Reminder Difficult to Understand?
Error Analysis of Synthetic Speech

- Synthetic versions worse (30% lower scores) for six medication names “Accumycin”, “Beclotor”, “Erytozole”, “Mevacycline”, “Pravaclor”, “Sulfacillin”

- Potential causes:
  - bad joins: Accumycin, Sulfacillin
  - short transitions: Beclotor, Erytozole, Mevacycline, Pravaclor

Discussion of stimuli later (with PRAAT TextGrid)
Making TTS Accessible to Older People

Pilot Study: Reminders in the Home
- Experiment Design
- Participants and Assessments

Results
- Effect of Stimuli
- Effect of Cognitive Ageing
- Effect of Auditory Ageing

Conclusion
- Summary
- Future Work
What is the Effect of Cognitive Ageing?

There is a clear difference between groups:

<table>
<thead>
<tr>
<th>Group</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>younger</td>
<td>38.9 ± 5.6</td>
</tr>
<tr>
<td>older</td>
<td>27.8 ± 8.6</td>
</tr>
</tbody>
</table>

But it only affects natural speech, contrary to expectations.

<table>
<thead>
<tr>
<th>Score</th>
<th>Correlation</th>
<th>95% Conf. Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural speech</td>
<td>$\rho=0.42$</td>
<td>[0.14, 0.64]</td>
</tr>
<tr>
<td>Synth. speech</td>
<td>$\rho=0.23$</td>
<td>[-0.07, 0.49]</td>
</tr>
</tbody>
</table>
Outline

1. Making TTS Accessible to Older People
2. Pilot Study: Reminders in the Home
   - Experiment Design
   - Participants and Assessments
3. Results
   - Effect of Stimuli
   - Effect of Cognitive Ageing
   - Effect of Auditory Ageing
4. Conclusion
   - Summary
   - Future Work
What is the Effect of Auditory Ageing?

Summary of Results

- **Hearing loss** affects how well people can understand synthetic speech.
- The factors that affect the intelligibility of natural speech and synthetic speech are **different**.
- We find these effects even with people who would pass normal screening tests (average of 0.5, 1, 2, and 4 kHz < 20dB).

Focus on results are presented for 35 participants (12 younger, 23 older) that would have passed screening.
Analysis Method

- Correlations between audiological measures and scores
- Separate measures for each ear, except for binaural gap detection
- Sub-thresholds for pure-tone audiometry data:
  - F2: 1, 2, 3 kHz
  - Trad: 0.5, 1, 2, 4 kHz
  - Normal: 0.25, 0.5, 1, 2, 3, 4, 6, 8 kHz
- EHF data: Average of all thresholds. If threshold could not be measured, value of maximum + 5dB used
Results

Strong Correlations ($p<0.005$)

Overall score / synthetic stimuli and

- F2 - left ear
- UHF - left and right ear
## Results

Weak Correlations ($p < 0.05$)

<table>
<thead>
<tr>
<th>Errors on . . .</th>
<th>Voice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Natural</td>
</tr>
<tr>
<td>Phonemes</td>
<td>PTA</td>
</tr>
<tr>
<td>Syllables</td>
<td>none</td>
</tr>
<tr>
<td>Words</td>
<td>Central</td>
</tr>
</tbody>
</table>

PTA: averaged thresholds  
Central: speech audiometry, gap detection
Outline

1. Making TTS Accessible to Older People
2. Pilot Study: Reminders in the Home
   - Experiment Design
   - Participants and Assessments
3. Results
   - Effect of Stimuli
   - Effect of Cognitive Ageing
   - Effect of Auditory Ageing
4. Conclusion
   - Summary
   - Future Work
Summary I: Audiological

- Hearing loss affects how well people understand synthetic speech, even for “normal subjects”
- The key frequencies are in the F2 region
- Overall condition of cochlea also important (EHF results)
Summary II: Methodological

- It is dangerous to extrapolate results on the intelligibility of human speech to synthetic speech and vice versa.
- Proper audiological evaluation (at least full PTA) pays off.
Extrapolation: Guidelines for Adapting TTS to Hearing Loss

- Use familiar words
- Avoid signal processing distortions
- Watch durations, in particular for transitions into and out of key phrases (result of post-hoc analysis).
  Easy hack: Put key information in separate intonational phrases and slow down slightly to bias system towards hyperarticulated units

All of these need to be tested experimentally on a range of systems
Making TTS Accessible to Older People

Pilot Study: Reminders in the Home
- Experiment Design
- Participants and Assessments

Results
- Effect of Stimuli
- Effect of Cognitive Ageing
- Effect of Auditory Ageing

Conclusion
- Summary
- Future Work
Future Work I: Validity

Methodological Validity:

- Compare against standard intelligibility testing methods, both audiological and “technological”
- Replicate with Festival and DECTalk
- Use shorter battery with better tests of central auditory processing

Ecological Validity: What makes TTS intelligible in background noise?

Guideline Validity: Test guidelines experimentally
Future Work II: Taking Research Further

Technological Improvements:
- Use hyperarticulated speech for key content
- Look at effect of glottal source model
- Extend work to hearing aid wearers

Understanding Relevant Factors:
- Investigate effect of temporal distortions
- Concise model of relevant audiological factors that are as orthogonal as possible
Questions?

- Further statistics in appendix
- References in paper

Sponsors: SPARC, MATCH
Further Statistics
Older People Have More Problems With Synthetic Speech

![Box plot showing score comparison between human and synthetic voices for older and younger groups.](image-url)
Higher F2-Threshold $\rightarrow$ Synthetic Voice Less Intelligible

$\rho = -0.53$, $p < 0.005$
Higher EHF Threshold $\rightarrow$ Synthetic Voice Less Intelligible

$\rho = -0.49$, $p < 0.005$
No Effect of PTA Thresholds on Perception of Human Speech

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Trad</th>
<th>F1</th>
<th>F2</th>
<th>EHF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Left Ear</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natural</td>
<td>-0.20</td>
<td>-0.16</td>
<td>-0.15</td>
<td>0.05</td>
</tr>
<tr>
<td>Synthetic</td>
<td><strong>-0.46</strong>*</td>
<td><strong>-0.53</strong>**</td>
<td><strong>-0.49</strong>**</td>
<td></td>
</tr>
<tr>
<td><strong>Right Ear</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natural</td>
<td>-0.14</td>
<td>-0.20</td>
<td>-0.11</td>
<td>-0.05</td>
</tr>
<tr>
<td>Synthetic</td>
<td>-0.36</td>
<td>-0.09</td>
<td><strong>-0.46</strong>*</td>
<td><strong>-0.50</strong>**</td>
</tr>
</tbody>
</table>
Left Ear Thresholds Stronger Predictors than Right Ear Thresholds

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Trad</th>
<th>F1</th>
<th>F2</th>
<th>EHF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Left Ear</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Synthetic</td>
<td>-0.43</td>
<td>-0.22</td>
<td>-0.45*</td>
<td>-0.30</td>
</tr>
<tr>
<td>Synthetic</td>
<td>-0.46*</td>
<td>-0.19</td>
<td>-0.53**</td>
<td>-0.49**</td>
</tr>
<tr>
<td><strong>Right Ear</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Synthetic</td>
<td>-0.33</td>
<td>-0.18</td>
<td>-0.37</td>
<td>-0.36</td>
</tr>
<tr>
<td>Synthetic</td>
<td>-0.36</td>
<td>-0.09</td>
<td>-0.46*</td>
<td>-0.50**</td>
</tr>
</tbody>
</table>