Voice-input voice-output communication aid (VIVOCA)

Simon Judge, Mark Hawley, Fabian Cardinaux, Peter O’Neil, Rebecca Palmer
Barnsley AT Team

- Assistive Technology (AT) team, covering 3 areas of S Yorkshire
- Assess for and provide a wide variety of AT
- Run training and provide support on AT
- Contribute to & run research and development projects...
Research Groups

- Barnsley District General Hospital Foundation Trust – R&D Department, AT Team
- Sheffield University – Computer Science Dept, Health and Related Sciences School
- Collaboration & Track record on AT projects.
- New group forming involving AT
- CAST group: Clinical Applications of Speech Technology
Dysarthria is the most common acquired speech disorder (170 per 100,000)

Many current communication aids (VOCAs) are slow and effortful to use

Dysarthric speech can be an effective control input to assistive technology
CAST Projects

NEAT
- STARDUST
  (Speech recognition for people with severe dysarthria)

SPECS
(Speech-enabled Control Systems)

Suite of products for older and disabled people

VIVOCA
(Voice-input voice-output communication aid)

HTD

NHS Innovation Hub
(Medipex)
Voice-input voice-output communication aid

Microphone

Speech Recogniser

‘Translation’ algorithm

Speech synthesiser
User-centred design & development

User requirements consultation → Concepts and early models

User consultation and trials ← Prototypes ← Final prototype

User trials

Clinical trial
# User and professional consultation

<table>
<thead>
<tr>
<th>Method</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOCA users and speech therapists</td>
<td>- Acceptable as a means of communication</td>
</tr>
<tr>
<td>Semi-structured interviews and focus groups</td>
<td>- Potential advantages over conventional VOCA</td>
</tr>
<tr>
<td>Thematic analysis</td>
<td>- Quicker</td>
</tr>
<tr>
<td></td>
<td>- Easier to use</td>
</tr>
<tr>
<td></td>
<td>- Increased communication and independence</td>
</tr>
<tr>
<td></td>
<td>- Useful where speed and intelligibility crucial</td>
</tr>
<tr>
<td></td>
<td>- Meeting new people</td>
</tr>
<tr>
<td></td>
<td>- Telephone</td>
</tr>
<tr>
<td></td>
<td>- Shopping</td>
</tr>
<tr>
<td></td>
<td>- Range of requirements for hardware and software</td>
</tr>
</tbody>
</table>
Speech recogniser for dysarthric speech

- Commercial speech recognisers do not work well for dysarthric speech
- User-centred approach – aim to make it work
- Speaker dependent recognition
- Vocabulary of discrete words tailored to speech capabilities of individual
- Closed loop between recogniser training and user training
Speech recogniser for dysarthric speech

Initial training data

Recogniser training

Additional training data

User training

Final recogniser models
Training: User Feedback

Chart Type: Concentric Circle

Score: 78%
Training: User Feedback

Chart Type: Concentric Circle

Score: 36%
## ‘Translation’ methods

<table>
<thead>
<tr>
<th>Small Number of inputs</th>
<th>Large vocabularies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5</td>
<td>~10</td>
</tr>
<tr>
<td>~10</td>
<td>~30</td>
</tr>
<tr>
<td>~30</td>
<td>~100</td>
</tr>
<tr>
<td>~100</td>
<td>~1000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Switch scanning</th>
<th>Mobile phone/T9</th>
<th>Spelling</th>
<th>AAC fixed overlay</th>
<th>Word for word translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morse code</td>
<td></td>
<td>Word prediction</td>
<td>AAC dynamic screen</td>
<td></td>
</tr>
</tbody>
</table>

**Coding**

**Direct ‘translation’**
Translation: input-output

**Input**
- Word
- Letter combination
- 2-3 word combination
- Word & letter combination

**Output**
- Word
- Word
- Phrase
- Phrase

"Want", "B", "R"  
Can I have a beer please?
Current speech synthesis: communication aids

- High quality voices available
- E.g.
  - DECTalk™ (Fonix) for American English
  - Acapela for British English
- Personalisation limited: age, gender, language
Personalisation

- Voice = identity
  - Gender
  - Age
  - Geographic background
  - Socio-economic background
  - Ethnic background
  - As that individual

- Maintains social relationships
- Maintains social closeness
- Sets group membership
VIVOCA: personalisation

- Sheffield/Barnsley user group
- Retain local accent
  - geographic identity
- Speaker database
  - Arctic database:
    - 593 + 20 sentences
- Professional local speakers
  - Ian McMillan
  - Christa Ackroyd
Concatenative synthesis

Speech recordings

Input data → Unit segmentation → Unit database

Text input

Unit selection → Concatenation + smoothing → Synthesised speech

Festvox: http://festvox.org/
Concatenative synthesis

- High quality
- Natural sounding
- Sounds like original speaker
- Need a lot of data (~600 sentences)
- Can be inconsistent
- Difficult to manipulate prosody
HMM synthesis

Model states

Time step

Generated observations

Speech output

yes

yes
HMM synthesis procedure

Speech recordings

Input data → Training → Synthesis

Speaker model

Text input

Synthesised speech

HTS http://hts.sp.nitech.ac.jp/
HMM synthesis

- Consistent
- Intelligible
- Needs relatively little input (~20 mins)
- Can be adapted with small amount of data (>5 sentences)
- Easier to manipulate
  - Buzzy quality
  - Less natural than concatenative
Synthesis: Future research

- Further personalisation for individuals with progressive speech disorders
  - Capturing the essence of a voice
- Voice banking
  - Before deterioration
- Adaptation using HMM synthesis
  - Before or during deterioration
Summary

- Voice in-Voice out device based on a PDA
- Currently under development
- Recognising and improving discrete dysarthric words
- Regionalised, possibly personalised, speech synthesis
VIVOCA Team

Mark Hawley, Pam Enderby, Phil Green, Stuart Cunningham, Peter O’Neill, Rebecca Palmer, Siddharth Sehgal, Fabien Cardinaux, Andre Coy, Simon Judge

www.barnsleyrd.nhs.uk

www.shef.ac.uk/cast

simon.judge@nhs.net
A Plug!!

- RAATE 2007
- 26^{th} and 27^{th} November 2007

www.raate.org.uk

(session on voice recognition!)