The NST 'homeService' application: recent system and experimental developments

Heidi Christensen, Iñigo Casanuevo, Stuart Cunningham, Phil Green, Thomas Hain
homeService: a challenging application for Natural Speech Technology

Putting state-of-the-art speech techniques, developed by the NST research team, to use in people's homes.

Exemplar Applications
- richer, more natural interaction

Learning & Adaptation
- long-term adaptation
- environmental modelling
- adapt ASR models
- who spoke when annotations

Natural Synthesis
- situation appropriate voices
- distant microphone ASR
- elderly and dysarthric speech ASR

Natural Transcription
- NST General meeting, Edinburgh 23. May 2013
Two sides of homeService

User study / System development
- online, ‘in-the-field’ recognition
- hardware
- software
- design of study
- ethics permission

Experimental Work
- offline, lab-based recognition
- using existing databases
- develop methods for handling elderly and dysarthric speech ASR

NST General meeting, Edinburgh 23. May 2013
homeService study

- Speech-enabled assistive technology (v1: environmental control, later versions with more natural speech interactions)

- For people who can’t or don’t like using remote controls, switches, keyboard etc.

- Study outline (as per our NHS ethics application):
  - 10 users (about 5 elderly and 5 physically disabled)
  - Longitudinal study; each user involved for 1-2 years
  - ‘Virtuous circle’ design where the system capabilities (functionality- and modelling-wise) are improved as new data is collected.
  - Permission to save all audio recorded whilst the system is being used
  - Plus special ‘open-mic’ sessions of about an hour. This will be un-annotated but unique, conversational data!
homeService system - hardware

- Off-the-shelf components
- Custom-built software
- Typical devices are TV, PC, radio, lamp, curtains
homeService system - software

**atHome** (v1) system controls peripherals (microcone, tablet, infrared transmitter), and sends audio back to the lab.

**atLab** system receives audio, connects to ASR server and has “intelligence” to decide next actions.
Recognition “in-the-cloud”

Benefits:

- ASR acoustic and language models can be updated remotely
- Each user will have different system version and setup
- Enables remote login and monitoring
- System state is logged; all data is saved with time stamps
homeService demonstration
The other side of homeService

User study / System development
- online, ‘in-the-field’ recognition
- hardware
- software
- design of study
- ethics permission

Experimental Work
- offline, lab-based recognition
- using existing databases
- develop training strategy for dysarthric and elderly speech ASR
homeService experimental work

RESEARCH QUESTIONS:

• How can we best use methodologies and data from main-stream, typical speech ASR?
  
  • Investigate the use of state-of-the-art training strategies for dysarthric speech
  
  • Investigate use of typical speech knowledge for dysarthric speech

• How do we best ‘tune’ an ASR system to the non-typical elements of a dysarthric speaker?
  
  • Investigate the automatic derivation of pronunciation dictionaries for dysarthric speech

• How do we -- for a given speaker -- find the best `operating point’ for a personalised, homeService type ASR system?
  
  • Investigate ways of setting up initial system: choice of vocabulary, enrolment data requirement, etc.
Using methodologies and data from typical speech ASR

<table>
<thead>
<tr>
<th>Models</th>
<th>Absolute Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical speech meeting models</td>
<td>22.4%</td>
</tr>
<tr>
<td>UAspeech speaker-independent (SI) models</td>
<td>29.7%</td>
</tr>
<tr>
<td>USspeech speaker-dependent (SD) models</td>
<td>50.9%</td>
</tr>
<tr>
<td>UAspeech SI models + MAP adaptation</td>
<td>54.1%</td>
</tr>
<tr>
<td>UAspeech SI-MAP + Dynamic Neural Net features (providing typical speech phone knowledge)</td>
<td>62.5%</td>
</tr>
</tbody>
</table>

UAspeech: Largest English database of dysarthric speech, 18hrs, 16 speakers
Word Accuracy: isolated word task, word loop grammar restricted to one word recognised per utterance.

WE FOUND:

- Using standard triphone, HMMs trained according to typical speech training strategies works
- Adding Maximum A Posteriori (MAP) adaptation helps
- Adding further typical speech knowledge (through the use of neural net based features) helps
Using methodologies and data from typical speech ASR

<table>
<thead>
<tr>
<th>Models</th>
<th>Speaker Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical speech meeting models</td>
<td>22.4% (0.6% - 70.2%)</td>
</tr>
<tr>
<td>UAspeech speaker-independent (SI) models</td>
<td>29.7% (3.2% - 68.5%)</td>
</tr>
<tr>
<td>USspeech speaker-dependent (SD) models</td>
<td>50.9% (9.0% - 85.6%)</td>
</tr>
<tr>
<td>UAspeech SI models + MAP adaptation</td>
<td>54.1% (8.3% - 89.6%)</td>
</tr>
<tr>
<td>UAspeech SI-MAP + Dynamic Neural Net features (providing typical speech phone knowledge)</td>
<td>62.5% (9.9% - 92.6%)</td>
</tr>
</tbody>
</table>

UAspeech: Largest, English database of dysarthric speech, 18hrs, 16 speakers
Word Accuracy: isolated word task, word loop grammar restricted to one word recognised per utterance.

WE FOUND:

- Using standard triphone, HMMs trained according to typical speech training strategies works
- Adding Maximum A Posteriori (MAP) adaptation helps
- Adding further typical speech knowledge (through the use of neural net based features) helps
Tuning an ASR system to the non-typical elements of a dysarthric speaker

WE FOUND:

• Customising the dictionary can improve performance
• Best results when modifying only words that tend to get mis-recognised for a speaker.
Finding the best `operating point’ for a particular speaker

WE FOUND:

- For this speaker, having about 10 examples of each word gets you a reasonable performance.
What next?

homeService system:

- First user to test system is imminent, followed by staggered start of remaining users over the next year
- First release of data to rest of NST
- Incorporation of NST-investigated, e.g., on environmental modelling, personalisation and adaptation into real systems.

more research questions:

- Elderly speech
- Long-term adaptation
- How to use un-annotated data (from the `open-mic’ sessions)
Thank you