Speech recognition in Alzheimer’s disease with personal assistive robots

Frank Rudzicz$^{1,2}$, Rosalie Wang$^1$, Momotaz Begum$^3$, Alex Mihailidis$^{4,1}$

$^1$ Toronto Rehabilitation Institute,
$^2$ Department of Computer Science, University of Toronto,
$^3$ Department of Computer Science, University of Massachusetts Lowell,
$^4$ Department of Occupational Science and Occupational Therapy, University of Toronto

Fifth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2014) 26 June 2014, Baltimore USA
Introduction

• Alzheimer’s disease (AD) is a progressive neuro-degenerative dementia characterized by declines in:
  • Cognitive ability (e.g., memory, visual-spatial reasoning),
  • Functional capacity (e.g., executive power), and
  • Social ability (e.g., linguistic abilities).

• Caregivers often assist individuals with AD, either at home or in long-term care facilities.
• >$100B are spent annually in the U.S. on caregiving for AD.
• As the population ages, the incidence of AD may double or triple in the next decade (Bharucha et al., 2009).

• Demographic crisis!
• ‘COACH’ automates support of daily tasks often assisted by human caregivers.
  • E.g., hand-washing, tooth-brushing.
  • Based on partially-observable Markov decision processes (POMDPs) and vision-only input.

• But what if the user does not want to spend their day in front of the sink?
ED the robot

Our goal is to implement two-way spoken dialogue in ED that can **identify** and **recover** from communication breakdowns.
Common features in dialogue in AD: Repetition, incomplete words, and paraphrasing (Guinn and Habash, 2012).

- Pauses, filler words, formulaic speech, and restarts were not.
  - Surprisingly, this seems to contradict Davis and Maclagan (2009), and Snover et al. (2004).

- Effects of AD on syntax remains controversial.
- Agrammatism could be due to memory deficits (Reilly et al., 2011).

Pakhomov et al. (2010) found pause-to-word and pronoun-to-noun ratios were discriminative of frontotemporal lobar degeneration.

Roark et al. (2011) found pause frequency and duration were indicative of mild cognitive impairment.
Data collection: tea for two

- Ten individuals (6 female) with AD recruited at Toronto Rehab.
  - Age: 77.8 years ($\sigma = 9.8$)
  - Education: 13.8 years ($\sigma = 2.7$)
  - MMSE: 20.8/30 ($\sigma = 5.5$)

- Three phases with different partners:
  - A **familiar** human-human dyad (during informed consent),
  - A human-robot dyad (during **tea-making**), and
  - An **unfamiliar** human-human dyad (during post-study interview).
Our data are very noisy. Signal-to-noise: –2.1 dB to 7.63 dB.
Clean speech typically 40 dB to 60 dB.
Can we do speech recognition in this environment accurately?

We assume that our recordings can be decomposed as:

$$y(t) = x(t) + d(t)$$
Noise reduction

- Subtraction with log-spectral amplitude estimator (LSAE)
  - Requires an annotated sample of the noise.
Noise reduction

Moderate

Severe
Speech recognition

• Semi-continuous hidden Markov model with 42-dimensional MFCC input (incl. $\delta$ and $\delta\delta$), $z$-scaled.

• Two trigram language models derived from English Gigaword (small: top 5000 words, large: top 64,000 words).

• Five speaker-independent acoustic models derived from WSJ over 100 speakers with 1, 2, 4, 8, and 16 Gaussians/state.

• Empirically adjust other parameters (e.g., beam width).
## Results

<table>
<thead>
<tr>
<th>Vocab.</th>
<th>Scenario</th>
<th>Noise reduction</th>
<th>AD (%)</th>
<th>Caregiver (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>Interview</td>
<td>None</td>
<td>25.1 (σ = 9.9)</td>
<td>28.8 (σ = 6.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSAE</td>
<td>40.9 (σ = 5.6)</td>
<td>40.2 (σ = 5.3)</td>
</tr>
<tr>
<td></td>
<td>In task</td>
<td>None</td>
<td>13.7 (σ = 3.7)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSAE</td>
<td>38.2 (σ = 6.3)</td>
<td>35.1 (σ = 11.2)</td>
</tr>
<tr>
<td>Large</td>
<td>Interview</td>
<td>None</td>
<td>23.7 (σ = 12.9)</td>
<td>27.0 (σ = 10.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSAE</td>
<td>38.2 (σ = 6.3)</td>
<td>35.1 (σ = 11.2)</td>
</tr>
<tr>
<td></td>
<td>In task</td>
<td>None</td>
<td>5.8 (σ = 3.7)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSAE</td>
<td>14.3 (σ = 12.8)</td>
<td>-</td>
</tr>
</tbody>
</table>

**Results**

- **Small Interview**
  - None: 25.1 (σ = 9.9), 28.8 (σ = 6.0)
  - LSAE: 40.9 (σ = 5.6), 40.2 (σ = 5.3)
  - *t*(58) = 3.9, *p* < 0.005

- **Large Interview**
  - None: 23.7 (σ = 12.9), 27.0 (σ = 10.0)
  - LSAE: 38.2 (σ = 6.3), 35.1 (σ = 11.2)
  - *t*(39) = 8.7, *p* < 0.0001
Accuracy and MMSE

- Despite the clear increasing trend in accuracy with MMSE, $n$-way ANOVA:
  
  $F_1 = 47.07$, $p = 0.164$
Communication strategies

• To be useful, ED needs to mimic some verbal techniques employed by caregivers.

• Caregivers are commonly trained to use communication strategies (Small et al., 2003), such as:
  • Using a relatively slow rate of speech,
  • Repeating misunderstood prompts verbatim,
  • Posing closed-ended questions (e.g., yes/no questions),
  • Simplifying the syntactic complexity of sentences,
  • Giving one question or one direction at a time, and
  • Using pronouns minimally.
How to identify breakdowns?

• **Trouble Indicating Behaviors (TIB)** (Watson, 1999).
  • Difficulties can be phonological, morpho/syntactic, semantic (e.g., lexical access), discourse (e.g., misunderstanding topic).
  • 7 seniors with AD use TIBs significantly more ($p < 0.005$) than matched controls (Watson, 1999).

• >33% of moderate AD dyads display related ‘**trouble-source repair**’ (Orange, Lubinsky, Higginbotham, 1996).
  • **Most common trouble:** discourse (e.g., inattention, working memory)
  • **Most common repair:** *wh*-questions and hypotheses (e.g., “Do you mean ...?”).
How to identify breakdowns?

- People with AD were much ($t(18) = -5.8, p < 0.0001$) more likely to exhibit **TIB 8 (lack of uptake)** with the robot...
How to identify breakdowns?

• ... people with AD were much more likely ($t(18) = -4.78$, $p < 0.0001$) to have **successful** interactions with a **robot** (18.1%) than with a non-familiar **human** (6.7%).
Ongoing work

• We can achieve up to 40% word accuracy in AD using standard acoustic/language models and noise reduction.
  • Accuracy depends on MMSE, but not significantly.
  • We are currently improving ASR by adapting vocabularies, acoustic and language models.

• Older adults with AD are very likely to ignore the robot, but when they don’t they have more fluid dialogues than with unfamiliar humans.

• Automatically identify TIBs from > 200 acoustic and lexical/syntactic features with an accuracy of TODO%.