Overview
Evaluation of Automatic Speech Recognition (ASR) systems using data collected from six deployed dialogue systems.

Speech recognizers tested
Cambridge HTK family
HVite (v3.4.1), HDecode & Julius (v4.1.2)
- Two sets of acoustic and language models:
  - Trained directly on TRAIN
  - Adapted with the WSJ training corpus
- Same acoustic and language models for all engines.

CMU Sphinx family
Sphinx 4 & Pocket Sphinx (v0.5)
- Language model trained directly on TRAIN
- Acoustic model adapted with the WSJ training corpus

Data
Five domains of human speech directed at one or more virtual characters:
- SGT Blackwell
  - A question-answering character who answers general questions about the Army, himself, and his technology.
  - "When did you join the Army?"
  - "How can you understand what I'm saying right now?"
- SGT Star
  - A question-answering character who talks about careers in the Army.
  - "Who are you?"
  - "Is the pay good in the Army?"
- Amani
  - A bargaining character used as a prototype for training soldiers to perform tactical questioning.
  - "Do you know where he lives?"
  - "I'll keep this a secret."

Report.
A training prototype that responds to military calls for fire in a virtual reality urban combat environment.
- "M T O kilo alpha four rounds target number alpha one out."
- "Shot out."

One additional domain of conversations between two human participants:
- IOTA
  - An extension of the Radiobots system with more varied types of calls for fire (including calls for air support).
  - "Roger where do you want me to go?"
  - "Roger contact on that east west road."

Corpus Size
Corpora divided into TRAIN (≈80%), DEV (≈10%) and TEST (≈10%)

Vocabulary
Vocabulary (TRAIN) vs. Out-of-Vocab. Items (DEV)

Results
Main result: Word Error Rate on TEST

<table>
<thead>
<tr>
<th></th>
<th>Non-Real-time</th>
<th>Real-time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HVite</td>
<td>HDecode</td>
</tr>
<tr>
<td>Blackwell</td>
<td>34</td>
<td>30</td>
</tr>
<tr>
<td>Star</td>
<td>34</td>
<td>35</td>
</tr>
<tr>
<td>Amani</td>
<td>36</td>
<td>41</td>
</tr>
<tr>
<td>SASSO</td>
<td>35</td>
<td>39</td>
</tr>
<tr>
<td>Radiobots</td>
<td>36</td>
<td>42</td>
</tr>
</tbody>
</table>

- The difference between recognizers is typically smaller than the differences between the domains.
- Among the HTK recognizers, HDecode had the best word error rates in general. However, HDecode does not run in real-time.

Word Error Rate on DEV

<table>
<thead>
<tr>
<th></th>
<th>Non-Real-time</th>
<th>Real-time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HVite</td>
<td>HDecode</td>
</tr>
<tr>
<td>Blackwell</td>
<td>34</td>
<td>30</td>
</tr>
<tr>
<td>Star</td>
<td>34</td>
<td>35</td>
</tr>
<tr>
<td>Amani</td>
<td>36</td>
<td>41</td>
</tr>
<tr>
<td>SASSO</td>
<td>35</td>
<td>39</td>
</tr>
<tr>
<td>Radiobots</td>
<td>36</td>
<td>42</td>
</tr>
</tbody>
</table>

- Between the two Sphinx recognizers, PocketSphinx (real-time) outperformed Sphinx4 (not real-time) in five of the six datasets.

Adaptation Affects Performance
- HTK decoders tested on DEV under two conditions:
  - Unadapted: specific data set was used for training.
  - Adapted: the domain-specific dataset was augmented with a larger WSJ dataset.
  - Adaptation was done for both language and acoustic models.

- Some decoders were tested with bigram and trigram language models.

Adaptation

<table>
<thead>
<tr>
<th></th>
<th>HVite</th>
<th>HDecode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train (TRN)</td>
<td>Dev (DEV)</td>
</tr>
<tr>
<td>Blackwell</td>
<td>34</td>
<td>30</td>
</tr>
<tr>
<td>Star</td>
<td>34</td>
<td>35</td>
</tr>
<tr>
<td>Amani</td>
<td>36</td>
<td>41</td>
</tr>
<tr>
<td>SASSO</td>
<td>35</td>
<td>39</td>
</tr>
<tr>
<td>Radiobots</td>
<td>36</td>
<td>42</td>
</tr>
</tbody>
</table>

- No single setting is best in all data sets.
- HVite is a better adaptation for Blackwell, but worse for IOTA and Amani.
- HDecode does best with adapted trigrams for most domains.

- Adaptation decreases decoding speed because the search space is widened.
- Enriched models could compensate for data sparsity.
- Using the additional WSJ dataset increases the size of models substantially.

- The effect of adaptation is typically smaller than the inherent differences between the domains.

Conclusion
- No single ASR engine is superior to all others in every domain.
- WER levels for dialogues coming from different systems vary from relatively low to very high in different domains.
- Performance of free off-the-shelf ASR engines is good enough for use in virtual human dialogue systems in some specific applications, but not others.
- Additional considerations are important for practical applications:
  - Design requirements, e.g. whether real-time performance is needed.
  - Development constraints, e.g. ease of integration with other system components.

- In future work we intend to examine the impact of speech recognizers on the performance of natural language understanding modules and on overall performance of dialogue systems.

Acknowledgments
The work described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

The bulk of the work was performed at ICT, where the first two authors were interns in the summer of 2009.

The first author acknowledges financial support from the Erasmus Mundus Program through scholarships for the European Masters Program in Language and Communication Technologies (LCT).