Priority Theme 3: Socially Aware Interactive Assistant

Joseph Mariani
CNRS-LIMSI & IMMI, France
Contributors: M. Pantic, H. Bourlard, A. Waibel, K. Jokinen, G. Riccardi, S. Renals, J. Mariani
META-NET Interactive Systems Vision Group, META-Council

META-Forum 2012 A Strategy for Multilingual Europe
Brussels, Belgium, June 20/21

Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME, CESAR, METANET4U, META-NORD (grant agreements no. 249119, 271022, 270893, 270899).
Socially Aware Interactive Assistants

- Multilingual Assistants which support Human Interaction
  - Human-Computer Interaction,
  - Human-Artificial Agent (robot) interaction
  - Computer-mediated Human Interaction

- Acting in various environments
  - Instrumented environments ((meeting) rooms, offices, apartments)
  - Instrumented open environments (streets, cities, transportation, roads)
  - Web, Virtual / Augmented worlds (incl. (serious) games)

- Personalized to user’s needs and environment

- Learning incrementally and individually from all sources and interactions
The Socially Aware Interactive Assistants can:

- Interact naturally with you, wherever you are, in any environment
- Interact naturally with your relatives, wherever they are
- Interact in any language and in any communication modality
- Adapt and personalize to individual communication abilities (handicap)
- Transcribe into text any fluent speech, pronounce fluently any text
- Self-Assess its performances and recover from errors
- Learn, personalize & forget through natural interaction
- Act on objects in instrumented spaces (rooms, apartments, streets)
- Assist in language training and education in general
- Provide a synthetic multimedia information analysis (knowledge)
- Recognize people’s identity, & their gender, accent, language, style, age
- Move, manipulate objects, touch people (Robots)
Global Abilities

- Interact naturally with Agents (terminals, ECA, robots, chatbots, humans, things)
  - In games, entertainment, education, communication, instrumented spaces, Call Centers, etc.
- Communicate everywhere
  - Mobile applications, Augmented Reality
- In society
  - Social networks and forums, Multiparty communication including several humans, several artificial agents/robots
- In a personalized way
  - Person, Gender, Style, Age
  - Accent, Language
Domain-specific Abilities

- Exhibit language proficiency
  - Speech-to-Speech Translation, Interpretation in meetings or in videoconferences,
  - Cross-lingual information access
- Overcome handicap obstacles
  - Crossmodal/crossmedia, Assistive applications, Sign Language
  - Adapted communication (cars, meetings)
- Refer to written support
  - Speech transcription, Subtitling
  - Reading machine
- Provide personalized training
  - Computer Aided Language Learning
  - Incl. dialects
  - Education, (self-)assessment
Tentative roadmap

- Where are we?
- Try to define time to deliver
  - Should reach “good enough” quality
- Socially Aware Interactive Assistants
  - Global Abilities
  - Domain-specific Abilities
    - Aid to Multilingualism
    - Aid to the “authors” (speech-text)
    - Aid to the handicapped
    - Aid to education and training
- Resources and Evaluation
Where are we?

- What are the performances offered by the technologies?
  - What are the performances needed by the application?
  - Automatic Speech Recognition
    - NIST ASR evaluation
    - Voice Command and Voice Dictation achieve performances close to humans (2-4% WER)
    - Radio/TV broadcast transcription (10% WER) doesn’t achieve human performances but sufficient for automatic indexing
    - Conversational speech and Meeting transcription performances are not sufficient (50% WER), especially for languages other than American English
  - Machine Translation
    - Euromatrix+
    - MT achieve performances getting close to human translators for few EU language pairs (Maltese-to-English) (Machine: 70% – Human: 80% BLEU)
    - Far behind for most EU language pairs
    - For Text translation: Speech translation?
Where are we?

- New US Babel program (2012-2016)
  - Funded by US IARPA
  - ASR for other languages than American English
  - 22 languages in many different language families
    - Afro-Asiatic, Niger-Congo, Sino-Tibetan, Austronesian, Dravidian, Altaic,…
  - 4 “surprise languages”: the goal is to be able at final (2016) to develop an ASR system for a new language within a week

- Apple SIRI
  - Intelligent personal assistant and knowledge navigator
  - Initially (Fall 2011): 4 languages / 6 varieties
  - Next (iOS 6): 9 languages / 20 varieties
  - Consider countries not languages: also includes cultural dimension
# Socially Aware Interactive Assistants

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Interacting naturally with agents (terminals, ECA, robots, things) (in games, entertainment, education, communication, instrumented spaces, Call Centers, etc.)</td>
<td>![Diagram](Q A H Ref)</td>
<td>![Diagram](D A Ref)</td>
<td>![Diagram](Ref A A)</td>
</tr>
</tbody>
</table>

19-20.06.12

META-FORUM 2012 - Brussels
# Roadmap (excerpt)

<table>
<thead>
<tr>
<th>Research Priorities</th>
<th>Targeted Breakthroughs in Technology Development</th>
</tr>
</thead>
</table>
| **Interacting naturally with agents**                    | **Provide usable human interface,**
| (terminals, ECA, robots, things)                          | Reliable speech recognition, natural and intelligible speech synthesis, limited understanding and dialog capabilities |
|                                                          | **Provide usable dialog interface,**
|                                                          | Context and dialog aware speech recognition and synthesis. Recognize and produce emotions, understanding capabilities, context aware dialog, using other sensors (GPS, RFID, cameras, etc.) |
|                                                          | **Provide multiparty (Human-Agents) interface,** multiple voices, mimicking, advanced understanding and advanced personalized dialog (indirect speech acts, incl. prosodics (lies, humor)) |

19-20.06.12

META-FORUM 2012 - Brussels
### Roadmap (excerpt)

<table>
<thead>
<tr>
<th>Research Priorities</th>
<th>Targeted Breakthroughs in Technology Development</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Using language but also other modalities, in parallel or together</strong></td>
<td>Multimodal interaction (speech, facial expression, gesture, body postures)</td>
</tr>
<tr>
<td><strong>Conscious of its performing capacities</strong></td>
<td>Confidence in hearing/understanding, interactively recovering from mistakes</td>
</tr>
</tbody>
</table>
# Roadmap (excerpt)

<table>
<thead>
<tr>
<th>Research Priorities</th>
<th>Targeted Breakthroughs in Technology Development</th>
</tr>
</thead>
</table>
| **Exhibiting multilingual proficiency** *(speech-to-speech translation, interpretation in meetings and videoconferences, crosslingual information access)* *(Theme 1)* | 2013-2014: Ensure availability or portability to major EU languages (30). Recognize which language is spoken. Multilingual access to multilingual information  
## Roadmap (excerpt)

<table>
<thead>
<tr>
<th>Research Priorities</th>
<th>Targeted Breakthroughs in Technology Development</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Resources (Theme 4)</strong></td>
<td><strong>2013-2014</strong></td>
</tr>
<tr>
<td><strong>Install infrastructure</strong></td>
<td>Install infrastructure</td>
</tr>
<tr>
<td><strong>Collection of multi-task</strong></td>
<td>Collection of multi-task</td>
</tr>
<tr>
<td><strong>benchmark data.</strong></td>
<td>benchmark data.</td>
</tr>
<tr>
<td><strong>Collaborative</strong></td>
<td>Collaborative</td>
</tr>
<tr>
<td><strong>production of</strong></td>
<td>production of</td>
</tr>
<tr>
<td><strong>semantically annotated</strong></td>
<td>semantically annotated</td>
</tr>
<tr>
<td><strong>data (multimodal).</strong></td>
<td>data (multimodal).</td>
</tr>
<tr>
<td><strong>Incremental</strong></td>
<td>Incremental</td>
</tr>
<tr>
<td><strong>production of dialog</strong></td>
<td>production of dialog</td>
</tr>
<tr>
<td><strong>data.</strong></td>
<td>data.</td>
</tr>
<tr>
<td><strong>In all EU languages</strong></td>
<td>In all EU languages</td>
</tr>
</tbody>
</table>
## Roadmap (excerpt)

<table>
<thead>
<tr>
<th>Research Priorities</th>
<th>Targeted Breakthroughs in Technology Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-task benchmark evaluation. Measures and protocols for automated speech synthesis, dialog systems and speech translation evaluation.</td>
<td>Measure of progress / Phase 1</td>
</tr>
<tr>
<td>For all EU languages.</td>
<td>More languages</td>
</tr>
</tbody>
</table>
Thank you.

office@meta-net.eu

http://www.meta-net.eu
http://www.facebook.com/META.Alliance