Interfacing with the Machine

Jay Desloge
SENS Corporation

Sumit Basu
Microsoft Research
They (We) Are Better Than We Think!

• Machine source separation, localization, and recognition are not as distant as they may seem.

• There are, in fact, already systems that achieve limited success in these areas.

• These machines provide many opportunities to investigate the interaction of machines with the human operator.
Consider: Hearing Aids

- Directional microphones can yield target-location (in front of wearer) intelligibility-weighted SNR improvements of up to 5-6 dB.
- Adaptive directional capability can yield higher SNR improvements (on the order of 8-12 dB).
- FM capability allows aid to receive signals from remote sources (TVs, remote microphones).
Consider: Tele/Video Conferencing

- Directional microphones used to identify and extract the sources from the environment. IW SNR improvements 5-6 dB on average.
- Active speaker is determined by microphone input.
- Voice-tracking capability can focus video camera on an active source within the environment. RMS loc. error < 10 deg.

(Polycom Soundpoint)
Consider: ASR State of the Art

<table>
<thead>
<tr>
<th>Type</th>
<th>Characteristics</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting Room (16kHz)</td>
<td>Task oriented, but includes true meetings collected in uncontrolled conditions</td>
<td>30% (head-mounted)</td>
</tr>
<tr>
<td></td>
<td>Far-talking, but also have close-talking (head-mounted) for comparison</td>
<td>50% (distant)</td>
</tr>
<tr>
<td>Switchboard (Telephone)</td>
<td>Close-talking, relatively free of noise; these are real people (with a slight bias toward females housewives and higher education), who don’t know each other and have some conversation on some topic. Real data, but instrumented Conditions</td>
<td>15%</td>
</tr>
<tr>
<td>Broadcast News</td>
<td>“Found data” (exists in nature, not artificially collected); Spoken by professional speakers; not read, but speakers know what they are going to say in advance, and possibly Practice</td>
<td>9%</td>
</tr>
<tr>
<td>WSJ (Dictation)</td>
<td>High-quality microphones, professional speakers, “Wall Street Journal” sentences (ie it’s a rich, but restricted domain)</td>
<td>3-8%</td>
</tr>
<tr>
<td>String of Digits</td>
<td>Easy task; no noise, close-talk</td>
<td>&lt;0.5%</td>
</tr>
</tbody>
</table>

From Patrick Nguyen (MSR)
Consider: Wireless Communication, GPS

• Wireless communication links can connect team members (e.g., military, firefighter, police) and can provide clean, separated signals for each source.

• GPS can provide accurate information about the location of each source.

• Efforts have already been made to present these sources to the team members in a logical manner (e.g., spatialized audio).
What We Will Talk About

Given that these and other possibilities for human-machine interaction already exist, it is important to study how the humans and machines can interact in a manner that achieves the best possible performance.

We will discuss:

- Machine enhancement of human capabilities (H+)
- Human enhancement of machine performance (M+)
- Design factors in human-machine interfaces
Machines Enhancing Human Capabilities (H+)

• Despite their limitations, machines can outdo what we do
H+: Going Beyond the Human Scale

• **Very large arrays:**
  – Localization for low-frequencies
  – Localization for impulsive/wideband sounds

H+: Augmenting Ears

• The strength of numbers:
  – As a localizer or recognizer, machines may be at about half human performance
  – With 100 sensors => 50 humans worth!!
  – But what good is a fractional human?

• State of the Art in General Sound Recognition
  – Speech detection
  – Everything else
H+: Multiplying Ears

...because there may be too many things to listen to...
H+: Multiplying Ears

...too many sounds in too many places...
H+: Distant Ears

…because we can’t be everywhere at once…
H+: Replacing Ears

...because we may have limited hearing capabilities...
H+: Augmenting Ears

...because we’re not always paying attention...
H+: The Sixth (Seventh, etc.) Sense

• We can apply existing techniques to frequency ranges/senses we don’t have
  – Ultrasound
  – Microwave
Humans Enhancing Machine Performance (M+)  

- Despite impressive machine computational capability, there are still certain tasks that the human can do faster and more reliably.
M+: What Do We Optimize?

• **Finding the right objective function is hard**
  – SNR vs. intelligibility
  – Listening comfort
  – Particularly true if a human will be listening to the output

• **Example: Hearing Aids**
M+: System focus

- Where are the sources?
M+: Environmental Conditions

- The human is often better at scene analysis
- Can drive system to optimize for varying conditions
  - Low Reverb? High Reverb?
  - Few, localized sources? Many sources?
M+: Calibration

- Some systems (e.g., conventional array processing) require knowledge of physical arrangement of microphones.

- Portable/body-mounted systems in particular must be configured and calibrated for proper operation.
Discussion and Teaser: Designing the Interactive System

- **Input from the user:**
  - How can we use direct manipulation and implicit manipulation to control the machine’s abilities?

- **Output to the user**
  - How do we decide what information is relevant to the user and how much they can handle?
  - How do we consolidate information into concise visuals/auralizations?
  - How can we display multiple auditory/visual streams to the user?