Interacting with Purpose (and Feeling!)

What neuropsychology and the performing arts can tell us about 'real' spoken language behaviour

Prof. Roger K. Moore
Chair of Spoken Language Processing
Dept. Computer Science, University of Sheffield, UK
(Visiting Prof., Dept. Phonetics, University College London)
(Visiting Prof., Bristol Robotics Laboratory)

Spoken Language Dialogue Systems

- Recent years have seen considerable progress
- Improvements have come from:
  - established standards (such as VoiceXML)
  - corpus-driven statistical modelling (e.g. POMDPs)
  - the introduction of paralinguistic behaviours (such as back-channeling and emotion)
- Progress has not come about as a result of deep insights into human spoken dialogue
- Spoken language dialogues can be
  - fragile (in 'real' conditions)
  - difficult (for users to inhabit)
  - expensive (to port to new applications / languages)
- There is concern in the R&D community about whether some applications will ever be realised
View of the R&D Community


2. Most telephone Interactive Voice Response systems accept speech input (and more than just digits)

<table>
<thead>
<tr>
<th>Year</th>
<th>2009</th>
<th>2003</th>
<th>1997</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>2015</td>
<td>2008</td>
<td>2002</td>
</tr>
<tr>
<td>SD</td>
<td>109</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Max</td>
<td>3220</td>
<td>2060</td>
<td>2020</td>
</tr>
</tbody>
</table>

© 2011 The University of Sheffield
5. Automatic airline reservation by voice over the telephone is the norm.

<table>
<thead>
<tr>
<th>Year</th>
<th>2009</th>
<th>2003</th>
<th>1997</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>2022</td>
<td>2010</td>
<td>2007</td>
</tr>
<tr>
<td>SD</td>
<td>10</td>
<td>10</td>
<td>57</td>
</tr>
<tr>
<td>Min</td>
<td>2001</td>
<td>2002</td>
<td>1999</td>
</tr>
<tr>
<td>Max</td>
<td>2080</td>
<td>2050</td>
<td>2500</td>
</tr>
</tbody>
</table>

17. Most routine business transactions take place between a human and a virtual personality (including an animated visual presence that looks like a human face).

<table>
<thead>
<tr>
<th>Year</th>
<th>2009</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>2060</td>
<td>2043</td>
</tr>
<tr>
<td>SD</td>
<td>118</td>
<td>66</td>
</tr>
<tr>
<td>Min</td>
<td>2012</td>
<td>1994</td>
</tr>
<tr>
<td>Max</td>
<td>3000</td>
<td>2500</td>
</tr>
</tbody>
</table>
An Insight into Current SL Dialogue
Usability of Dialogue Systems

Graph courtesy of Mike Phillips (CEO, Mobeus Corporation)

Usability

Like a Human

??

Add NL/Dialog

Structured Dialog

Flexibility

“Spoken language is the most sophisticated behaviour of the most complex organism in the known universe”
(Dawkins, 1991 + Gopnik et al, 2001)

So, what is ‘behaviour’, and how should we model it?
Q: Why does a living organism do anything?
A: Survival

(Notes: Some don’t do much!)

Two competing views …
– stimulus-response (‘behaviourism’)
– needs, drives and goals


Behaviour

- An organism’s behaviour is conditioned on a set of basic needs and goals
- ‘Appraisal mechanisms’ assess any given situation with regard to an individual’s needs and goals
- The outcome of such an appraisal process is ‘emotion’
- Emotion drives action (to meet needs)


The Purpose of Emotion

“These observations suggest that appreciation of the system’s own performance (entropy of success) and regulating the decision making strategy may indeed be one of the main functions of the emotional system in the brain.”

Regulatory/Control Systems

Negative feedback provides stability

**Perceptual Control Theory (PCT)**


**Homeostasis / Autopoiesis**

An Affective Model of Behaviour

**Intention**
**Attention**
**Consequence**

**Emotion** (valence)

**Motivation/Effort** (arousal)

**Motor Command**

**Action**

**FORWARD PATH**

**FEEDBACK PATH**

Interacting Control Systems

**System**
**System’s needs**
The system’s needs should be: “to satisfy the user’s needs”

**User**
**User’s needs**

Emergent behaviour = accommodation and entrainment (but also conflict)

Emulation Mechanisms

- Negative feedback control works OK as long as the loop delays are not too high, but neural delays can be quite significant in living systems.
- Delays can be overcome using an ‘emulation’ mechanism (pseudo-closed-loop control).


Inhibition of motor commands facilitates mental imagery and planning.

Emulator mitigates slow proprioceptive feedback.

Emulating Other(s)

Recognising other’s external behaviour.

Modelling the ‘surface behaviour’ of another agent.
Emulate $\Rightarrow$ Predict $\Rightarrow$ Understand

- Simulations *(forward models)* generate top-down expectations/predictions
- The emulation of an organism’s abilities can be recruited to emulate the behaviour of others for:
  - predicting their behaviour, mimicry, learning
- The *(emergent)* consequences would be:
  - an ability to interpret the behaviour of other(s) by referencing self
  - an ability to estimate the ‘beliefs, desires and intentions’ of other(s)
  - i.e. action understanding
- This is hypothesised to be the basis of ‘empathy’ and ‘Theory of Mind’ *(Baron-Cohen, 1997)*
- Possible mechanism: ‘mirror neurons’ *(Rizzolatti et al, 1996)*


Empathy

Recognising other(s) affective state(s)

empathy

Modelling the ‘internal states’ of another agent
Theory of Mind

Recognising other’s beliefs & intentions

Modelling the ‘perspective’ of another agent

Cooperative Behaviour

- Organisms cooperate in order to improve the outcomes of predation and procreation
- Cooperation requires the synchronisation of behaviour between different individuals
- Synchronisation can only be achieved by the communication of information
- Communication between organisms is more efficient if it can exploit empathetic/ToM relationships between individuals
- Such communication paves the way for the particulate structure of language (Abler, 1989)
Language

Minimum effort / maximum entropy exploitation of the ‘perspective’ of another agent

Influencing other’s external behaviour & internal states

Cognitive Robots

robot-world interaction

robot-robot interaction

human-robot interaction
Language-based HRI

Both agents require ‘Theory of Mind’ (of each other)

PreSenCE: Predictive Sensorimotor Control & Emulation

PreSenCE: Predictive Sensorimotor Control & Emulation

S:n

feeling

S:i

intention

S:m

action

S:E(U:n)

S:E(U:i)

S:E(U:m)

S:E(U:E(S:i))

S:E(U:E(S:m))

S:E(U:m)

S:E(U:E(S:i))

S:E(U:i)

S:E(U:n)

motivation

feeling

sensitivity

attention

interpretation

action

needs

noise, distortion, reaction, disturbance

System’s emulation of User’s emulation of System

System’s interpretation of User behaviour


PreSenCE: Predictive Sensorimotor Control & Emulation

sensitivity

S:n

feeling

S:i

intention

S:m

action

S:E(U:n)

S:E(U:i)

S:E(U:m)

S:E(U:E(S:i))

S:E(U:E(S:m))

S:E(U:m)

S:E(U:E(S:i))

S:E(U:i)

S:E(U:n)

motivation

feeling

sensitivity

attention

interpretation

action

needs

noise, distortion, reaction, disturbance

System’s emulation of User’s emulation of System

System’s interpretation of User behaviour

**Implications for SLI**

- Spoken language needs to be modelled as a cognitively-driven phenomenon, not a peripheral behaviour.
- Ultimately, speech perception, speech generation and conversational interaction *cannot* be modelled independently.
- Speech needs to be modelled as a cognitive behaviour that is conditioned on a **communicative context** where:
  - the talker has in mind the *needs* of the listener(s)
  - the listener has in mind the *intentions* of the talker(s)
- Need to abandon the ‘speech chain’ in favour of the ‘**communicative loop**’

---

Implications for SLP

• A new model of speech generation/synthesis that:
  – selects its characteristics appropriate to the needs of the listener
  – monitors the effect of its own output
  – modifies its behaviour according to its internal model of the listener

• A new model of speech recognition/understanding that:
  – uses a forward/generative model based on an internal emulation of the intentions of the speaker
  – adapts its forward/generative model to the voice of the speaker based on knowledge of its own voice

• A new model of dialogue that:
  – is driven by the need to satisfy the user(s)’ needs
  – uses emotion to appraise its success

• No contemporary SLP systems exploit such sensorimotor integration, overlap & control

Example: Reactive Speech Synthesis

Example: Optimising User ‘Happiness’


Example: Short vs. Long Loops

Problem with Language-based HRI

Only one agent has ‘Theory of Mind’
*(the other has no mind!)*

The human doesn’t know the robot doesn’t have a mind

Solution #1: *Fake it!*
The ‘Uncanny Valley’


Solution #2: Be ‘Appropriate’

- We should be developing robots that are clearly robots, not facsimiles of human beings
- This has consequences for how they should look, sound and feel

Implications for SLI

- Recognising the needs and intentions of human beings is clearly an almost impossible task given...
  - the complexity and sophistication of human behaviour
  - our primitive understanding of how to configure a cognitive architecture (such as PreSenCE)

- In the short term, the theory underpinning PreSenCE leaves only one possible line of approach...
  - manage users’ expectations such that their behaviours fall automatically within the capabilities of the autonomous agent

- In practice, this suggests that a system should exhibit visual, vocal and interactional 'affordances' that make it apparent to a user just what it can and, more importantly, what it cannot do

- Inspiration on 'appropriate' characterisation can be gained from the performing arts

An ‘Appropriate’ Voice?


Social Engagement with Robots and Agents

<table>
<thead>
<tr>
<th></th>
<th>Dislocation not perceived</th>
<th>Dislocation perceived</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jolly voice</td>
<td>16 (69.6%)</td>
<td>1 (40.4%)</td>
</tr>
<tr>
<td>Smile voice</td>
<td>7 (20.1%)</td>
<td>16 (69.6%)</td>
</tr>
</tbody>
</table>

Summary & Conclusion

- R&D progress suggests that the field is moving in a positive direction.

- However, it is accumulating a battery of successful short-term engineering solutions rather than developing an underlying long-term theory of vocal interaction.

- This talk attempted to address this issue by:
  - drawing inspiration from research fields that are quite distinct from spoken language technology (such as psychology, neuroscience and the performing arts);
  - providing some useful insights into potential generic principles of human (and even animal) behaviour.

- A common theme emerged that focused on:
  - the need to model beliefs, desires, actions and consequences (BDIAC) within a layered control-feedback architecture;
  - the importance of ‘appropriate’ characterisation of advanced communicative agents (such as robots).
and finally …

A system with appropriate needs and behaviours!

Thank You

Any questions?

http://www.dcs.shef.ac.uk/~roger
Acknowledgements

- **ACORNS**
  - Acquisition of Communication and Recognition Skills

- **COMPANIONS**
  - Intelligent, Persistent, Personalised Multimodal Interfaces to the Internet

- **CREST**
  - The Creative Speech Technology Network

- **SERA**
  - Social Engagement with Robots and Agents

- **S2S**
  - Sound to Sense

- **SCALE**
  - Speech Communication with Adaptive Learning

Relevant Research at USFD

Vocal Interactivity Laboratory

vocal interactivity in and between humans, animals and robots