On Speaker-Listener-Environment Coupling

Implications for Computational Models of Spoken Language

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 Lab.)

EU-FP7-EASEL

Rich History of Technological Development

Marconi 'SR128' (1982)
Radio Rex (1922)
Dragon 'Naturally Speaking' (1997)
Apple’s "Siri" (2011)
Voice dictation on a Smartphone (2007)
Rich History of Technological Development

Apple’s “Siri” (2011)

Speech-to-Speech Translation

Baidu Announces Breakthrough In Speech Recognition, Claiming To Top Google And Apple
Past, Present & Future

Command and Control Systems

Dictation Systems

Interactive Voice Response (IVR) Systems

Voice-Enabled Personal Assistants

Embodied Conversational Agents (ECAs)

Autonomous Social Agents

The ‘State-of-the-Art’

• There is steady year-on-year progress

• Improvements come from:
  – increase in available computer power
  – corpus-driven statistical modelling
  – public benchmark testing

• Progress has not come about as a result of deep insights into human spoken language

• Spoken language technology is
  – fragile (in ‘real’ conditions)
  – expensive (to port to new applications / languages)

• Performance appears to be reaching an asymptote that is well short of human abilities
  – 20-50% word error rate on conversational speech
The ‘State-of-the-Art’

• There is steady year-on-year progress.
• Improvements come from:
  – increase in available computer power
  – corpus-driven statistical modelling
  – public benchmark testing

• Progress has not come about as a result of deep insights into human spoken language.

• Spoken language technology is fragile (in ‘real’ conditions) and expensive (to port to new applications / languages).

• Performance appears to be reaching an asymptote that is well short of human abilities – 20-50% word error rate on conversational speech.

‘Traditional’ SLP Architecture

So, what’s wrong with this? (if anything)
‘Traditional’ SLP Architecture

**Teleological Behaviour**

- The behaviour of *(intelligent)* living systems is intentional!
- This does not mean that an organism ‘knows’ what it is doing!
- It simply means that an organism has preferred states, and that actions are selected in order to achieve those states
- This places a focus, not on actions, but on the consequences of actions
- This, in turn, leads to very interesting forms of coupling between …
  - an agent and its environment
  - an agent and another agent

---

Communicating Intentions

- Signalling involves physical/mental effort
- Large effort creates clear signals but uses more energy (and vice versa)
- The ‘target’ is a perception not a signal
- So optimisation is over competing perceptions not competing signals
- The intention is sufficient contrast at the pragmatic level (leading to suitable compensations at the semantic, syntactic, lexical, phonemic, phonetic and acoustic levels)
- The obstacles are ...
  - alternative interpretations (internal)
  - competing signals (external)

“I … do … not … know”
“I do not know”
“I don’t know”
“dunno”
“dunno”  [əə]


Motivation

- Desired consequences will only be achieved if an agent expends sufficient physical/mental effort
- The same is true for interpretation
- Sometimes large movements are necessary due to the need to overcome an obstacle in the environment
- However, living systems have evolved to minimise effort
- So the effort involved in behaviour is traded against the effectiveness of the end result
- Successful outcomes thus depend on the motivation, strength and knowledge of the agent
Feedback

- The structural coupling of an agent with its environment (including other agents) implies feedback
- Feedback is a regulatory process
- Feedback facilitates ...
  - the management of energy and entropy
  - the maintenance of stability
  - the comparison of achievements against intentions

“feedback ... is the central and determining factor in all observed behavior”


Evidence for Such Behaviour

- People naturally tend to speak louder/differently in noise (Lombard, 1911)
- Caregivers talk differently to children (Fernald, 1985)
- Speakers actively control articulatory effort (Lindblom, 1990)
- Users talk differently to machines (Moore & Morris, 1992)
- Being able to hear your own voice has a profound effect on speaking (as evidenced by the need for sidetone on a telephone)
- Hearing-impaired individuals can have great difficulty maintaining clear pronunciations (or level control)
- Delayed auditory feedback causes stuttering-like behaviour
- People with speaking difficulties (e.g. caused by cerebral palsy) report that it takes immense effort to produce even the simplest utterance
- Altered auditory feedback evokes compensations (Munhall et al, 2009; MacDonald et al, 2011)
Consequences for SLP

• Need modelling paradigms that are able to accommodate such dependencies

• Emphasises the importance of forward (generative) models

• Communicative obstacles are overcome using …
  – sufficient effort
  – feedback

• Communicative effort is related to …
  – the fidelity of the models
  – the depth of the searches
PreSenCE
Predictive Sensorimotor Control and Emulation


PreSenCE

S: needs → S: i → S: m → action

S: E: (m)

S: E: (m)

S: E: (i)

S: E: (n)

S: n

sensitivity

feeling

intention

action

Primary route for System’s motor behaviour

Emulation of System’s possible motor actions

- noise, disturbance

Emulation of System’s possible motor actions
PreSenCE

System's emulation of User's emulation of System

System's interpretation of User behaviour

action

noise, distortion, read

action

noise, distortion, read

S:E(U:n) S:E(U:i) S:E(U:m)

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

S:E(U:m)

S:E(U:m)

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

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)

S:E(U:m)

S:E(U:i)
PreSenCE

Affective signals drive the control loop(s)

Agent Interacting with Agent

Agents exploiting knowledge of themselves to manipulate/interpret each other

PreSenCE Related Research

• **ACORNS**
  – Acquisition of Communication and Recognition Skills

• **SERA**
  – Social Engagement with Robots and Agents

• **S2S**
  – Sound to Sense

• **SCALE**
  – Speech Communication with Adaptive Learning

• **COMPANIONS**
  – Intelligent, Persistent, Personalised Multimodal Interfaces to the Internet

Speech Energetics

‘AnTon’ – Animatronic Tongue


Reactive Speech Synthesis

Computational H&H

VPC: Vowel Production Control

CPC: Consonant Production Control


C2H: Experimental Setup

- HTS standard voice
  - British male voice
  - ~77,000 context-dependent models

- Trained using synthesised speech:
  - 2800 sentences synthesised with phone control sequences forced to have low-contrastive competitors
  - the most likely acoustic model for all phones is selected, even for those unseen in the original voice

- MLLR (Maximum Likelihood Linear Regression) transformation on models

Mauro Nicolao

VPC Results

<table>
<thead>
<tr>
<th>Feature</th>
<th>HYPO</th>
<th>NORM</th>
<th>HYPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Word Duration (s)</td>
<td>0.27</td>
<td>0.32</td>
<td>0.36</td>
</tr>
<tr>
<td>Mean Sentence Dur. (s)</td>
<td>2.98</td>
<td>3.50</td>
<td>3.91</td>
</tr>
<tr>
<td>Pause Duration (s)</td>
<td>0.13</td>
<td>0.15</td>
<td>0.17</td>
</tr>
<tr>
<td>LTAS 1-3 (dB SPL)</td>
<td>33.6</td>
<td>36.2</td>
<td>41.1</td>
</tr>
<tr>
<td>Spectral Tilt (dB/dec)</td>
<td>-6.2</td>
<td>-5.8</td>
<td>-4.7</td>
</tr>
<tr>
<td>Spectral CoG (Hz)</td>
<td>712</td>
<td>821</td>
<td>1024</td>
</tr>
<tr>
<td>F0 (Hz)</td>
<td>172.6</td>
<td>174.1</td>
<td>174.7</td>
</tr>
<tr>
<td>F0 range (Hz)</td>
<td>146-185</td>
<td>151-183</td>
<td>145-190</td>
</tr>
<tr>
<td>F1F2 area (Hz²)</td>
<td>1014</td>
<td>29021</td>
<td>70509</td>
</tr>
</tbody>
</table>

### CPC Results

<table>
<thead>
<tr>
<th></th>
<th>HYPO</th>
<th>NORM</th>
<th>HYPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Word Duration (s)</td>
<td>0.31</td>
<td>0.32</td>
<td>0.33</td>
</tr>
<tr>
<td>Mean Sentence Dur. (s)</td>
<td>3.43</td>
<td>3.50</td>
<td>3.60</td>
</tr>
<tr>
<td>Pause Duration (s)</td>
<td>0.14</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td>LTAS 1-3 (dB SPL)</td>
<td>35.4</td>
<td>36.2</td>
<td>38.4</td>
</tr>
<tr>
<td>Spectral Tilt (dB/dec)</td>
<td>-6.1</td>
<td>-5.8</td>
<td>-5.1</td>
</tr>
<tr>
<td>Spectral CoG (Hz)</td>
<td>547</td>
<td>821</td>
<td>1156</td>
</tr>
<tr>
<td>F0 (Hz)</td>
<td>174.1</td>
<td>174.1</td>
<td>173.4</td>
</tr>
<tr>
<td>F0 range (Hz)</td>
<td>144-185</td>
<td>151-183</td>
<td>150-184</td>
</tr>
<tr>
<td>F1F2 area (Hz²)</td>
<td>41824</td>
<td>29021</td>
<td>56103</td>
</tr>
</tbody>
</table>


### Effect on Vowel Space

\[ \alpha = 0.8 \]
Effect on Intelligibility

\[ \alpha = 0.8 \]

Example Speech: English Male

<table>
<thead>
<tr>
<th>Type of noise</th>
<th>HYPO</th>
<th>NORM</th>
<th>HYPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Shaped Noise (SNR = 1 dB)</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Competing Talker (SNR = -7 dB)</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Clean</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
</tbody>
</table>

“The box was thrown beside the parked truck”
### Example Speech: Italian Female

<table>
<thead>
<tr>
<th>Type of noise</th>
<th>HYPO</th>
<th>NORM</th>
<th>HYPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car Noise (SNR = -4 dB)</td>
<td>blue</td>
<td>grey</td>
<td>red</td>
</tr>
<tr>
<td>Babble Noise (SNR = -4 dB)</td>
<td>blue</td>
<td>grey</td>
<td>red</td>
</tr>
<tr>
<td>Competing Talkers (SNR = -4 dB)</td>
<td>blue</td>
<td>grey</td>
<td>red</td>
</tr>
<tr>
<td>Clean</td>
<td>blue</td>
<td>grey</td>
<td>red</td>
</tr>
</tbody>
</table>

*“Ti è mai successo di rimanere senza fiato?”*

---

### Example Speech: Italian Male

<table>
<thead>
<tr>
<th>Type of noise</th>
<th>HYPO</th>
<th>NORM</th>
<th>HYPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car Noise (SNR = -4 dB)</td>
<td>blue</td>
<td>grey</td>
<td>red</td>
</tr>
<tr>
<td>Babble Noise (SNR = -4 dB)</td>
<td>blue</td>
<td>grey</td>
<td>red</td>
</tr>
<tr>
<td>Competing Talkers (SNR = -4 dB)</td>
<td>blue</td>
<td>grey</td>
<td>red</td>
</tr>
<tr>
<td>Clean</td>
<td>blue</td>
<td>grey</td>
<td>red</td>
</tr>
</tbody>
</table>

---

And Finally …

Would we study walking by suspending someone in the air and asking them to walk?

No? So why do we put people in a recording booth and ask them to speak?

In both cases the subject is obliged to imagine a crucial conditioning aspect of their behaviour.

*An appropriate interactive experimental methodology is the key to future progress.*

---

‘Traditional’ Architecture

![Diagram of 'Traditional' Architecture]

---

Intentional Agent

User Model

Self Model

Environment

Multimodal Grounding

 Interpretation

Prediction

Action
The field of spoken language processing (SLP) typically treats speech as a stimulus-response process, hence there is strong interest in the SLP community in using the latest machine learning techniques to estimate the assumed static transforms.

This is especially true at the present time as evidenced by the huge growth in research using deep neural nets. However, in reality, speech is not a static process - rather it is a sophisticated joint behaviour resulting from actively managed dynamic coupling between speakers, listeners and their respective environments.

Multiple layers of feedback control play a crucial role in maintaining the necessary communicative stability, and this means that there are significant dependencies that are overlooked in contemporary SLP approaches.

This talk will address these issues in the wider context of intentional behaviour, and will give an insight into the implications of such a perspective for the next generation of computational models for spoken language processing.