Research Challenges in the Automation of Spoken Language Interaction

Prof. Roger K Moore
Chair of Spoken Language Processing
Speech and Hearing Research Group (SPandH)
Dept. Computer Science, Sheffield University, UK

Motivation

Some Good News 😊 …

– Spoken Language Interaction (SLI) plays a fundamental role in almost all human activities

– There is a burgeoning market for existing spoken language technology

– There are good prospects for automated SLI if it can deliver a clear benefit
Motivation

Some Bad News 😊 …

– Spoken language is a highly evolved social communication system
– Attempts to automate SLI raise significant research challenges at the forefront of human knowledge

“Spoken language interaction is the most sophisticated behaviour of the most complex organism in the known universe!”

Overview of Talk

• Markets and Applications
• Performance Modelling
• Prospects for the Future
• Research Challenges
• Overcoming Fragility
• Aspiring to a Universal Capability
• Summary and Conclusions
Markets & Applications

- Interactive vs non-interactive
  - conversational human-machine interface (A)
  - monitoring speech communications (B)

- On-line vs off-line
  - ‘live’ speech processing (B)
  - speech data mining (C)
Markets & Applications

“Radiology report number 5 6 3 dated 19 November 1998”

Medicine

“Engage ILS”

Transport

“... yours sincerely etc. End memo. Please mail by tonight.”

Office

Progress

Figure by permission of Prof. Sadaoki Furui, TIT, Japan.
Which Way to Go from Here?

“The industry has yet to bridge the gap between what people want and what it can deliver.”

“Reducing the ASR error rate remains the greatest challenge.”

“After sixty years of concentrated research and development in speech synthesis and text-to-speech (TTS), our gadgets, gizmos, executive toys and appliances still do not speak to us intelligently.”

Use formal methods to select R&D directions

Challenge #1
What Technology is Needed?

- Practical solutions can only be derived in response to particular applications requirements
- The successful assessment of future prospects depends on a careful analysis of the key market drivers
- This process is known formally as **Technology Roadmapping**

Technology Roadmapping (TRM)

*"Needs-driven technology planning process"

Pioneered by Motorola in the 1980s
The TRM Process

Market

Product

Technology

Time

Product features

Market drivers

Analysis Grids

Technical solutions

Market

Technology

Product

Analysis

Grids


Current & Future Spoken Language Technology

Technical Features

Technical Capabilities

Application Capabilities

Technical Requirements

Application Requirements

Application Benefits

technical assessment

application assessment

’suitability’

technical factors

application factors

The ‘Capability Profile’

Market Requirements (Pull)

Product Performance Envelope

Technology Capabilities (Push)

Technology Migration

FEATURES

improve

benchmark

extend

optimise

combine

port

FEATURES

FEATURES

FEATURES

+
The Lessons from TRM

Success or failure of automated SLI depends on …
– the needs of the marketplace
– it being a viable and affordable alternative to other market solutions

… and some alternatives aren’t that good

One SLI Alternative …
Avoid the costs of assessing automated SLI empirically

Challenge #2

Performance Modelling

• TRM can be used to reverse engineer the performance requirements necessary to meet the stated goals

• What is really needed is a model-based approach to assessing the merits of alternative interactive technologies

• For example, is there a way of figuring out whether state-of-the-art ASR would create viable applications on a PDA without incurring the expense of actually building a system to find out?
Alternative Data Entry Methods

Speaking:

- LVCSR: large-vocabulary continuous speech recognition
  “The cat sat on the mat”

- ASR of spelled words
  “t”, “h”, “e”, “space”, “c”, “a”, “t”, … etc.

- ASR using the ICAO phonetic alphabet
  “tango”, “hotel”, “echo”, “space”, “charlie”, etc.

Typing, tapping, keying & writing:

- typing on a conventional QWERTY keyboard
- soft typing (tapping on a PDA using a stylus)
- multi press (pressing keys on a mobile phone keypad)
- T9® ‘text completion’ on a mobile phone keypad
- handwriting recognition (normal font)
A Model for Data Entry

Definitions

- entry rate (#wpm - words per minute)
- throughput (#cwpm after all errors have been corrected)

\[ T = \frac{60 \times R}{60 + (R \times E \times C)} \]

throughput (cwpm)

entry rate (wpm)

word error rate (%)

correction time (secs)

Calibrating the Model

- Model set up to reflect typical:
  - entry rates
  - error rates
  - correction times
- Data derived from the literature and measured from human subjects
  - e.g. average time to correct an ASR error:
    - 29.1 secs (voice correction)
    - 13.2 secs (multimodal correction)
Typical ‘Expert’ Entry Rates

<table>
<thead>
<tr>
<th>INPUT MODE</th>
<th>ENTRY RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>QWERTY typing</td>
<td>150 wpm</td>
</tr>
<tr>
<td>Dictating to ASR</td>
<td>107 wpm</td>
</tr>
<tr>
<td>T9®</td>
<td>46 wpm</td>
</tr>
<tr>
<td>Soft typing</td>
<td>43 wpm</td>
</tr>
<tr>
<td>Spoken Alphabet</td>
<td>30 wpm</td>
</tr>
<tr>
<td>Multi-press</td>
<td>27 wpm</td>
</tr>
<tr>
<td>ICAO Alphabet</td>
<td>17 wpm</td>
</tr>
<tr>
<td>Handwriting</td>
<td>16 wpm</td>
</tr>
</tbody>
</table>

ASR Configurations Investigated

- **SD LVCSR VC**: speaker-dependent large vocabulary continuous speech recognition with voice-based correction
- **SD LVCSR MMC**: speaker-dependent large vocabulary continuous speech recognition with multi-modal correction
- **SI LVCSR VC**: speaker-independent large vocabulary continuous speech recognition with voice-based correction
- **SI LVCSR MMC**: speaker-independent large vocabulary continuous speech recognition with multi-modal correction
- **SI Ortho VC**: speaker-independent recognition of spelled words using the orthographic alphabet with voice-based correction
- **SI ICAO VC**: speaker-independent of spelled words using the phonetic alphabet with voice-based correction
Results

- QWERTY keyboard is fastest
- Handwriting recognition is slowest
- Best 'hands-free' configuration
- Users' 'ideal' configuration

Input Method

Performance Modelling

- Performance modelling can also be used to predict the levels of performance necessary to outperform alternative interface technologies

- In this case, it was determined that the ASR error rates needed to be reduced by a **factor of five** in order to be truly competitive
Future Market Drivers

• Demand for cost-saving through automation
• Growth in market for mobile devices
• Legislation
• Demand for personalised services
• Growth in network-based information services, messaging and edutainment
• 3G connectivity

AMBIENT INTELLIGENCE

“The future generation of technologies in which computers and networks will be integrated into the everyday environment, rendering accessible a multitude of services and applications through easy-to-use human interfaces.”
Future Market Drivers

A PC will have the computational power of the human brain by 2019 and will be equivalent to 1000 human brains by 2029

+ Growth in bioengineering

= 

Creation of automated system with human-like characteristics

+ Replacement of human faculties with automated (prosthetic) processes

Early 2000s

• Translating telephones allow two people across the globe to speak to each other even if they do not speak the same language

• Telephones are answered by an intelligent answering machine that converses with the calling party to determine the nature and priority of the call
• The majority of text is created using continuous speech recognition
• Ubiquitous language user interfaces
• Most routine business transactions take place between a human and an animated visual presence that looks like a human face
• Pocket-sized reading machines for the visually impaired
• Translating telephones commonly used for many language pairs

- Three-dimensional virtual reality displays and auditory lenses used routinely for communication with other persons, the Web and virtual reality
- Most interaction with computing is through gestures and two-way natural-language spoken communication
- The vast majority of transactions include a simulated person
2029

- Implants are used to provide input and output between the human user and the world-wide computer network
- Neural implants available to enhance visual and auditory perception and interpretation, memory and reasoning
- Automated agents are learning on their own
- The majority of communications involving a human is between a human and a machine

IVR

SD: 10 (4)
Max: 2060 (2020)

Home Automation

Mean: 2016 (2010)
SD: 15 (12)
Max: 2100 (2100)

In-Car

Mean: 2016 (2008)
SD: 13 (8)
Max: 2075 (2050)
Routine Business Transactions

Mean: 2041 (2009)
SD: 66
Min: 1994
Max: 2500

Translating Telephones

Mean: 2057 (early 2000s)
SD: 116
Min: 2000
Max: 3000
Human-Computer Interaction

Mean: 2069 (2019)
SD: 225
Min: 2004
Max: 3827

Predict the future quantitatively

Challenge #3
Performance Extrapolation

- In ASR, the last 15 years have seen a consistent annual relative error-rate reduction of ~10% (~0.7 dB HENR)

- If this continues, ASR will be almost as accurate as the human ear in 10-60 years:
  - transcription of read newspaper text by 2010
  - transcription of freestyle speech by 2017
  - recognition of digit strings by 2052
  - recognition of alphabet letters by 2060
Performance Extrapolation

- 2 year-old
- 10 year-old
- 80 year-old
- >70 lifetimes

Word Error Rate (%) vs. Hours

Supervised
Unsupervised
Unsupervised (reduced LM training)

Usable

Break out of the data-driven stranglehold of current approaches

Challenge #4

Specific Research Objectives

• Pattern characterisation for prediction and generalisation:
  – not models of data
  – but models of relationships between data

• Richer structural inventory to accommodate:
  – simultaneous asynchronous behaviour
  – dynamic patterning
  – decomposed & conditioned dependencies (individuality)
  – acoustic, phonetic, linguistic & cognitive effort/distinctions

• Communicative and contrastive language models

• Situated and embodied environments for acquisition, evolution and learning

• Integration with emerging theories of neuro-cognitive function, e.g. in the new field of Cognitive Informatics
Cognitive Informatics

Emerging *transdisciplinary* field in the cognitive and information sciences

- Autonomous computing
- Intellectual foundations of informatics
- Information models of the brain
- Informatics foundations of software engineering
- Expressive mathematics
- Internal information processing mechanisms
- Software agent systems
- Ergonomics
- Informatics laws of software
- Knowledge representation
- Neural computation

Reduce the *fragility* of current automated SLI

Challenge #5
Speech is Robust …

SLI R&D Challenges

**ASR**
- Accept input that is:
  - spontaneous
  - emotional
  - whispered
  - accented
  - disfluent
  - interrupted
  - contaminated
  - from the elderly
  - from the young
  - OOV rich

**TTS**
- Deliver output that is:
  - communicative
  - intelligible
  - understandable
  - acceptable
  - appropriate
  - expressive
  - personalised
  - sympathetic
  - reactive

**Dialogue**
- Provide interaction that is:
  - usable
  - effective
  - efficient
  - engaging
  - cost-effective
  - enjoyable
  - multi-modal
  - portable

… and **adaptive**
Towards Universal SLI

- Application-independence
- User-independence
- Environment-independence
- Rapid-adaptation
- Learning

Summary

- Spoken language interaction plays a fundamental role in almost all human activities
- Automated SLI offers the potential to reduce the high costs involved in the provision of customer services
- This raises significant research challenges at the forefront of human knowledge
Conclusion

- Future prospects depend on overcoming:
  - the *empirical* approach to anticipating system behaviour
  - the *fragility* of current approaches

- Long-term prospects will be best served by aspiring to more universal capabilities that are *independent* of particular applications, environments and users

We Need to be Able to Do This …
Thankyou

Any questions ? …