ISCA Distinguished Lecture

Beyond Siri

Towards the Next Generation of Talking and Listening Machines

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.)

The University Of Sheffield.

SPANDH

international speech communication association

promoting international speech communication, science and technology

- Started in 1999 by combining …
  - ESCA (*European Speech Communication Association*)
  - ICSLP (*International Conference of Spoken Language Processing*)
- Purpose:
  - to promote Speech Communication Science and Technology, both in the industrial and academic areas
  - covering all the aspects of Speech Communication (*acoustics, phonetics, phonology, linguistics, natural language processing, artificial intelligence, cognitive science, signal processing, pattern recognition, etc.*)
- ISCA offers a wide range of services …
  - INTERSPEECH conference
  - ISCA workshops
  - SIGs (*special interest groups*)
  - Distinguished Lectures

The University Of Sheffield.

SPANDH

international speech communication association

promoting international speech communication, science and technology

ISCA Objectives:

- to stimulate scientific research and education,
- to organize conferences, courses and workshops,
- to publish, and to promote publication of scientific works,
- to promote the exchange of scientific views in the field of speech communication,
- to encourage the study of different languages,
- to collaborate with all related associations,
- to investigate industrial applications of research results,
- and, more generally, to promote relations between public and private, and between science and technology.

http://www.isca-speech.org

ISCA Distinguished Lecture

Beyond Siri

Towards the Next Generation of Talking and Listening Machines

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.)

Rich History of Technological Development

Von Kempelen's talking machine (1791)

Radio Rex (1922)

Parametric Artificial Talker (1953)

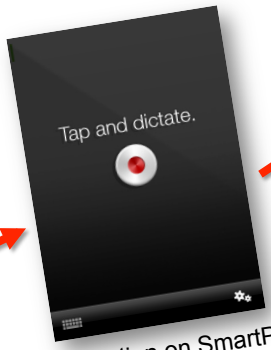
Speak'n'Spell (1983)

Interactive Talking Doll (1987)

Rich History of Technological Development

Marconi 'SR128' (1982)

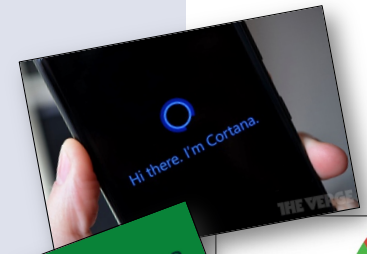
Dragon 'Naturally Speaking' (1997)



Voice dictation on SmartPhone (2007)

Apple's "Siri" (2011)

Rich History of Technological Development



Speech-to-Speech Translation

Apple's "Siri" (2011)

© 2015 The University of Sheffield

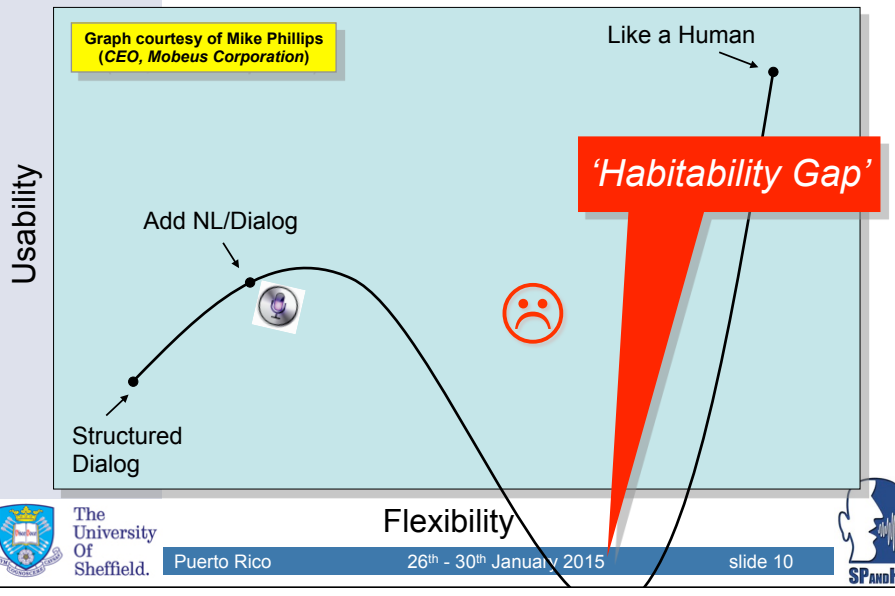
Rich History of Technological Development

How many 'real' users?

Apple's "Siri" (2011)

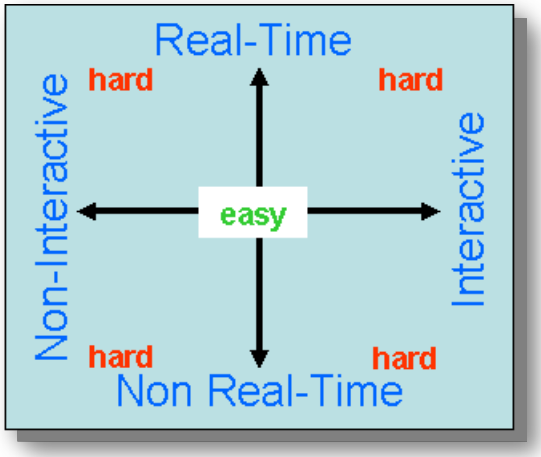
Still Some Way to Go?

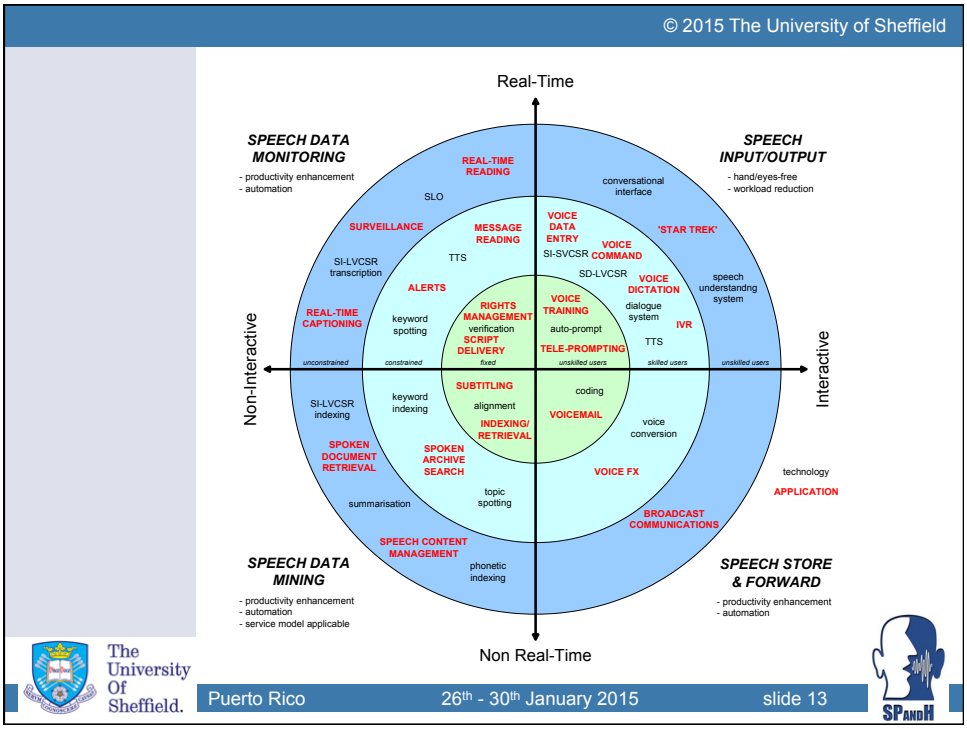
Still Some Way to Go?

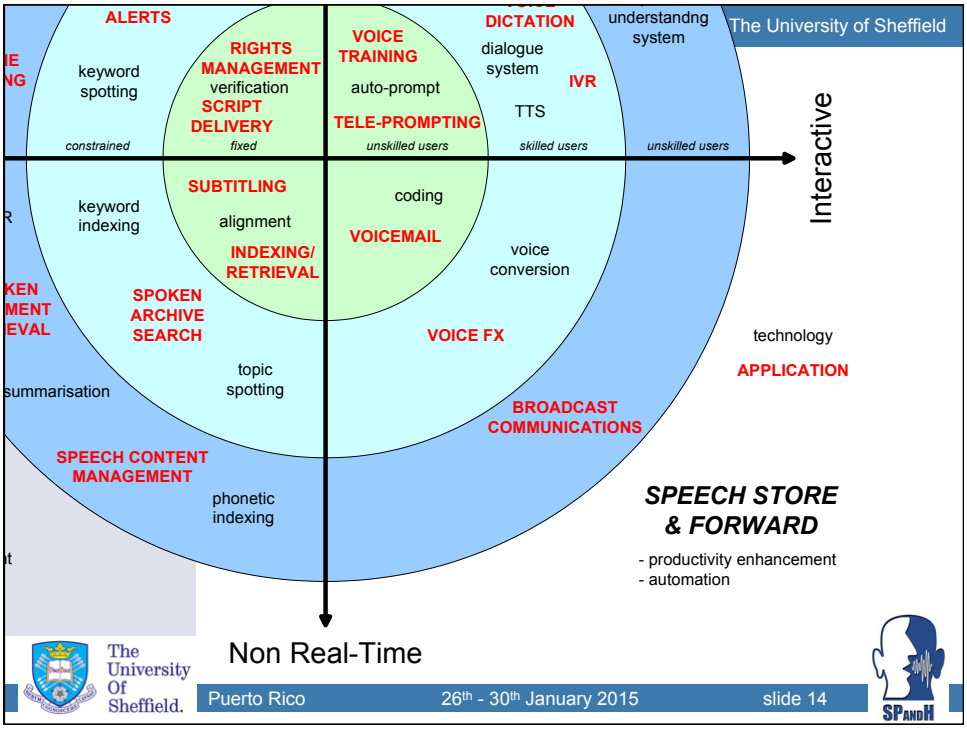


Still Some Way to Go?

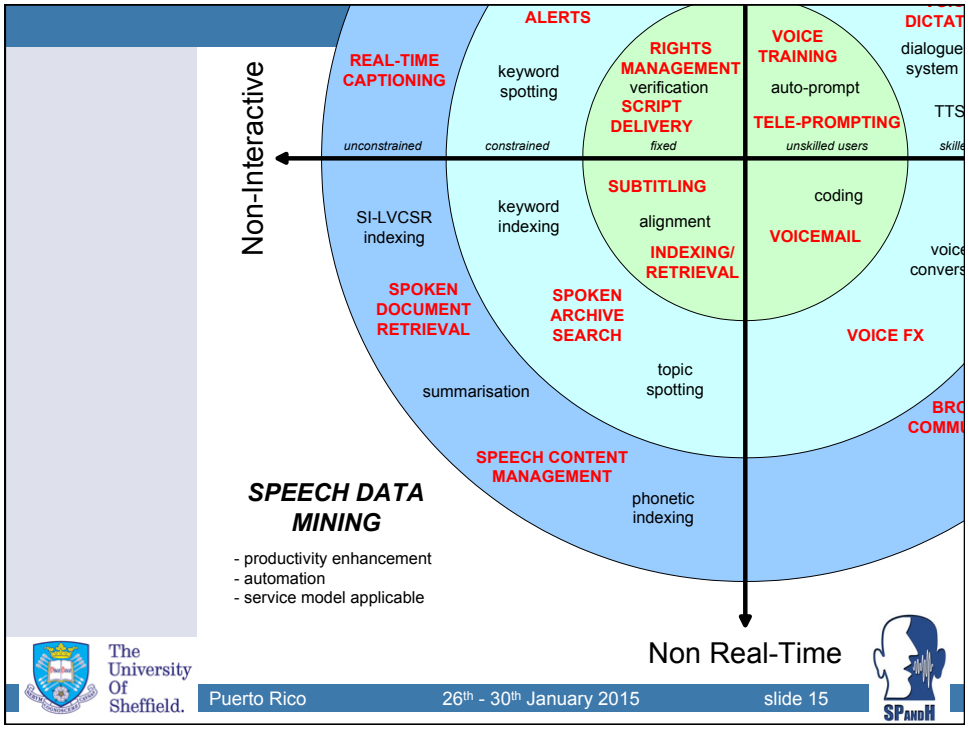
Taxonomy of SLP Applications

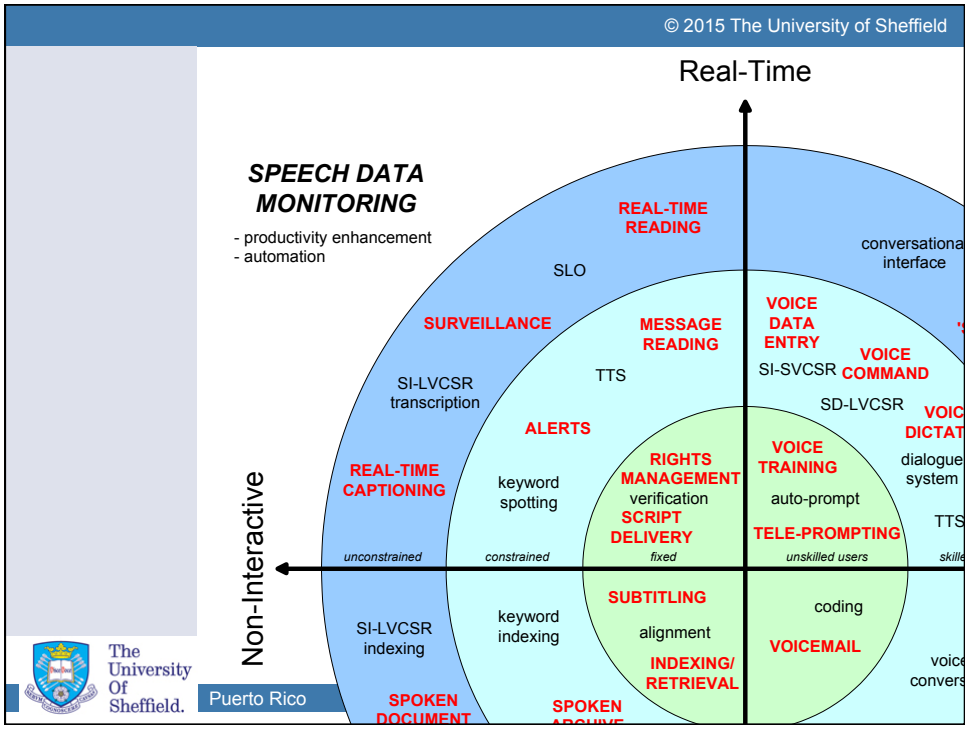


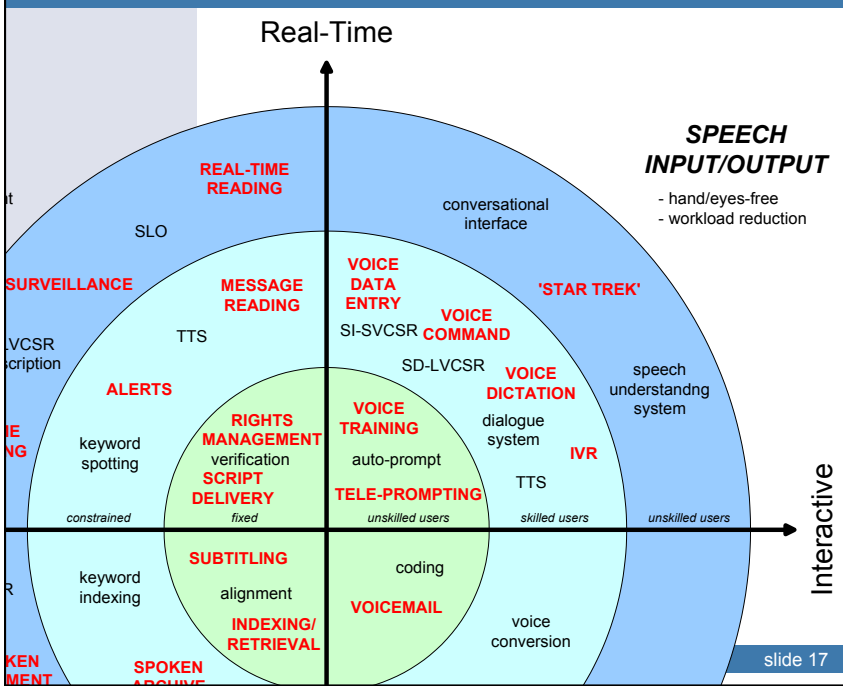




The University Of Sheffield.
Puerto Rico 26th - 30th January 2015 slide 14
SPandH


ALERTS
REAL-TIME CAPTIONING
keyword spotting
RIGHTS MANAGEMENT
verification
SCRIPT DELIVERY
VOICE TRAINING
auto-prompt
TELE-PROMPTING
dialogue system
TTS
unconstrained
constrained
fixed
unskilled users
Non-Interactive
SI-LVCSR indexing
keyword indexing
SUBTITLING
alignment
INDEXING/ RETRIEVAL
coding
VOICEMAIL
SPOKEN DOCUMENT RETRIEVAL
SPOKEN ARCHIVE SEARCH
VOICE FX
summarisation
topic spotting
SPEECH CONTENT MANAGEMENT
phonetic indexing
SPEECH DATA MINING
- productivity enhancement
- automation
- service model applicable
Non Real-Time
The University Of Sheffield.
Puerto Rico
26th - 30th January 2015
slide 15


© 2015 The University of Sheffield
Real-Time
SPEECH DATA MONITORING
- productivity enhancement
- automation
REAL-TIME READING
SLO
conversational interface
SURVEILLANCE
MESSAGE READING
TTS
VOICE DATA ENTRY
SI-SVCSR
VOICE COMMAND
SD-LVCSR
SI-LVCSR transcription
ALERTS
REAL-TIME CAPTIONING
keyword spotting
RIGHTS MANAGEMENT
verification
SCRIPT DELIVERY
VOICE TRAINING
auto-prompt
TELE-PROMPTING
dialogue system
TTS
unconstrained
constrained
fixed
unskilled users
Non-Interactive
SI-LVCSR indexing
keyword indexing
SUBTITLING
alignment
INDEXING/ RETRIEVAL
coding
VOICEMAIL
SPOKEN DOCUMENT
The University Of Sheffield.
Puerto Rico



Real-Time Interactive Applications

Command and Control Systems

Dictation Systems

Interactive Voice Response (*IVR*) Systems

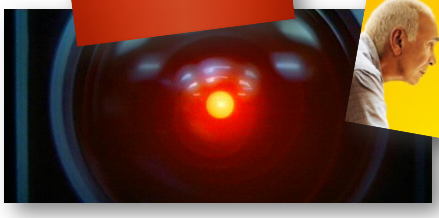
Voice-Enabled Personal Assistants

Embodied Conversational Agents (*ECAs*)

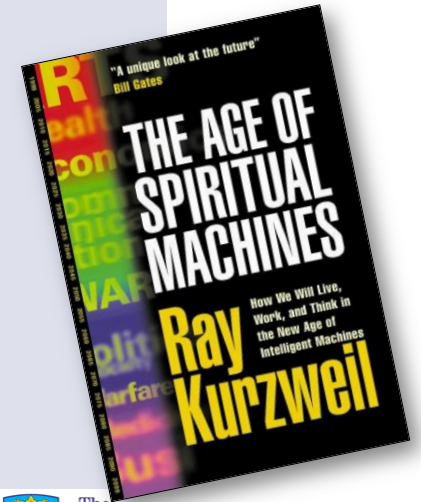
Autonomous Social Agents

The University Of Sheffield.

Autonomous Social Agents?



Autonomous Social Agents?



2019

"People are beginning to have relationships with automated personalities."

2029

"The majority of communications involving a human is between a human and a machine."

Autonomous Social Agents?

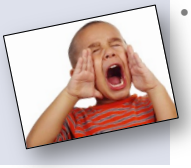
Beyond Siri

- Beyond speech
- Beyond words
- Beyond meaning
- Beyond communication
- Beyond dialogue
- Beyond one-off interactions

Beyond Speech

- Spoken language has evolved as part of a multimodal complex of interactive behaviours involving …
  - overall appearance
  - body posture
  - facial expressions
  - eye gaze
  - gestures and pointing

- Communicative information is distributed across these channels in a coordinated and coherent manner
- Speech is optimised according to the physical and temporal context of the interaction, e.g. …
  - words are chosen to clarify potentially obscure communicative points
  - the speed, loudness and clarity of speech are all adapted to the characteristics of the environment in which it is produced



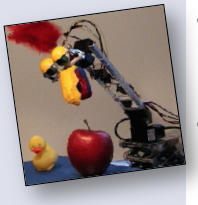
Beyond Speech

- The behavior of healthy living systems is mostly coordinated and coherent
- Any deviation from such consistency may be interpreted as physical or mental illness
- Unfortunately, this is exactly the situation that applies to many autonomous agents
- Mismatched capabilities can lead to confusion (*or even repulsion*) on the part of a user
- It is thus important that an autonomous social agent should be coherent in …
  - what it looks like
  - what it sounds like
  - what it says
  - how it behaves

Beyond Words

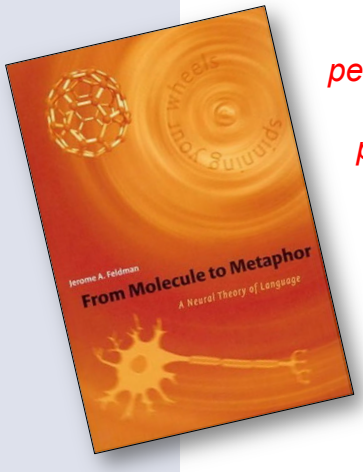
- The real communicative target is not words but meanings
- Automatic speech recognition is effectively a solved problem
- Figuring out why someone has said what you think they might have said (*and what you're supposed to do about it*) is an open challenge
- Likewise, it is relatively straightforward to configure a speech synthesiser to read out a defined sentence with a selected voice-type and prescribed prosodic contour
- What is more challenging is determine what to say, when to say it, and how such choices are manifest in the way in which an utterance is to be spoken

Beyond Words



- Unlike disembodied agents such as Siri, robots are instantiated as physical entities in the real world
- This means that a robot's behaviors are necessarily grounded in (*and constrained by*) the characteristics of its environment
- It is hypothesed that …
  - the meaning of language is grounded in bodily experience (*rather than in a prescribed ontology of logical forms*)
  - understanding is mediated by the use of metaphor as a mechanism for generalisation
- These ideas are supported by neurological evidence, driven largely by the discovery of so-called 'mirror neurons'

Beyond Words



"Understanding language about perceiving and moving involves much of the same neural circuitry as do perceiving and moving themselves."

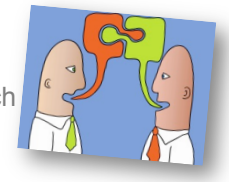
"The embodied neural approach to language suggests that the complex neural circuitry that supports grasping is the core meaning of the word."

Feldman, J. A. (2008). *From Molecules to Metaphor: A Neural Theory of Language*. Bradford Books.

Beyond Meaning

- Living systems are complex agents whose behaviours are determined by their drives, needs, beliefs and intentions
- An autonomous social agent needs to model such variables in order to interact successfully
- This requires the ability to interpret and express the paralinguistic phenomena which arise from the interaction
- E.g. if two agents have *convergent* intentions, then …
  - the effort required to perform a joint task may be shared
  - leading to the expression of satisfaction and pleasure
- Whereas, if two agents have *divergent* intentions, then …
  - the efforts required to perform a joint task may be vastly increased
  - leading to conflict between the participants and displays of displeasure

Beyond Communication

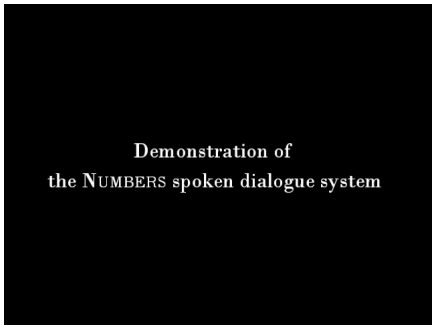


- Interaction between social species is much richer than simply exchanging messages using some shared code
- Behaviours are crafted to support continuous, coordinated interactions within/between social groupings
- Information is *not* just passed from one individual to another; available communication channels are exploited to manipulate the behaviour of others for cooperative/ competitive ends
- Also, information is not packaged into discrete nuggets; behaviours are aligned to provide continuous, adaptive coupling between individuals
- Artificial systems need to be able to model the co-active dynamic coupling between talking-listeners and listening-talkers

Beyond Dialogue

The traditional notion of strict turn-taking between a user and a spoken language dialogue system is giving way to a more fluid interaction based on partial hypotheses and incremental processing

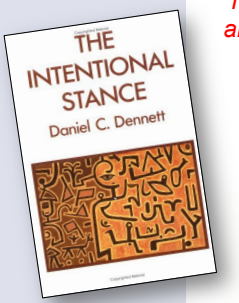
Schlangen, D., & Skantze, G. (2009). A general, abstract model of incremental dialogue processing. *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*. Athens, Greece.



Beyond Dialogue

- Interaction is grounded in the social relationships that exist between individuals …
  - the social status of the participants
  - the dominance relations between one individual and another
  - the trust that individuals put in each other
- These relations act as priors, i.e. they influence the way in which people talk
- Hence, an autonomous social agent *cannot* be viewed as a neutral partner; its perceived social status/believability will strongly influence the way in which users attempt to interact with it
- Failure to establish such relations at an early stage in a dialogue/interaction could lead to user confusion and the collapse of the interaction
- People typically employ small-talk to establish such relations

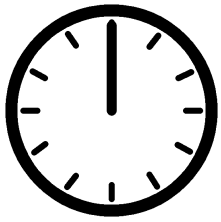
Beyond Dialogue



Dennett, D. (1989). *The Intentional Stance*. MIT Press.

'Design' Stance

"Things do what they are supposed to do"



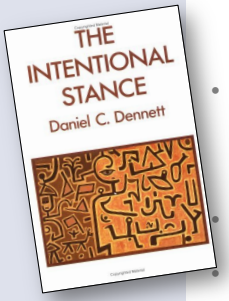
'Intentional' Stance

"Things do what they do on purpose"

'Physical' Stance

"Things obey the laws of physics and do what they do"

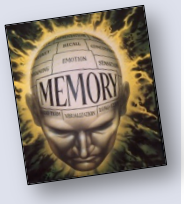
Beyond Dialogue



Dennett, D. (1989). *The Intentional Stance*. MIT Press.

- If a user takes a design stance to an object or device, then any unexpected behavior is taken to indicate that the device is broken and interaction should be abandoned
- If a user takes an intentional stance, then any unexpected behavior is taken as evidence that there are hidden motivations/goals that need to be determined (*and perhaps changed*)
- Users assign a 'Theory of Mind' to the agent/robot
- Interaction is unlikely to proceed if the agent is unable to explain its hidden mental states adequately
- The easiest way for an agent to be perceived as intentional, is for it to be intentional, i.e. to have its own internal needs and goals driving its behavior

Beyond One-Off Interactions



- People usually have a considerable prior history of interaction
- They retain person/context-specific memories of previous conversations, and are able to draw on this in order to interact efficiently
- Most human-robot interactions are short-term, with little or no memory (*in the robot*) of previous encounters
- The benefits of providing an autonomous social agent with a long-term memory are …
  - the facilitation of personalised conversational interaction
  - the opportunity to consolidate information from memory (*in order to be able to generalise to novel situations with known users or to novel users in known situations*)

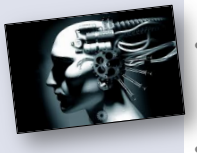
Overall Message

It is time to view spoken language as …

- *not* a faculty that is independent of other sensorimotor channels
- *not* a peripheral behaviour that is independent of core cognitive processes
- *not* an activity that is independent of the real-world context in which it takes place
- *not* a one-off exchange with no prior history

A Glimpse of the Future?

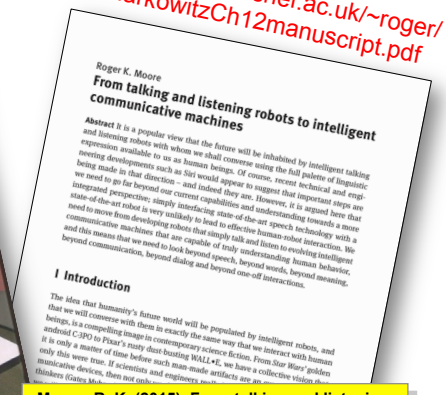
And finally … *Beyond Human?*



- A machine might one day be able to process speech better than a human being
- More interesting is the possibility of a machine doing something that a human being simply cannot do
- E.g. a huge disadvantage of a living system is its sensors and actuators are only able to function within a local area
- A machine's sensors and actuators may be distributed widely – throughout a home, across a city or around the planet (*i.e. an intelligent communicative machine's eyes, ears and mouths can literally be anywhere and everywhere*)
- It could have longer long-term memory, it could share all information amongst all agents and it could hold multiple conversations at the same time

Where to find out more …

http://www.dcs.shef.ac.uk/~roger/MarkowitzCh12manuscript.pdf



Roger K. Moore

From talking and listening robots to intelligent communicative machines

Abstract It is a popular view that the future will be inhabited by intelligent talking and listening robots with whom we shall converse using the full palette of linguistic expression available to us as human beings. Of course, recent technical and engineering developments such as Siri would appear to suggest that important steps are being made in that direction – and indeed they are. However, it is argued here that we need to go far beyond our current capabilities and understanding towards a more integrated perspective; simply interfacing state-of-the-art speech technology with a state-of-the-art robot is very unlikely to lead to effective human-robot interaction. We need to move from developing robots that simply talk and listen to evolving intelligent communicative machines that are capable of truly understanding human behavior, and this means that we need to look beyond speech, beyond words, beyond meaning, beyond communication, beyond dialog and beyond one-off interactions.

1 Introduction

The idea that humanity's future world will be populated by intelligent robots, and that we will converse with them in exactly the same way that we interact with human beings, is a compelling image in contemporary science fiction. From *Star Wars*' golden android C-3PO to Pixar's rusty chat-boasting WALL•E, we have a collective vision that it is only a matter of time before such man-made artifacts are no only this were true. If scientists and engineers municative devices, then not only thinkers (Gates, Mu

Moore, R. K. (2015). From talking and listening robots to intelligent communicative machines. In J. Markowitz (Ed.), *Robots That Talk and Listen* (pp. 317–335). Boston, MA: De Gruyter.

Over the past thirty years, the field of spoken language processing has made impressive progress from simple laboratory demonstrations to mainstream consumer products.

However, the limited abilities of commercial applications such as Siri highlight the fact that there is still some way to go in creating Autonomous Social Agents that are truly capable of conversing effectively with their human counterparts in real-world situations.

What seems to be missing is an overarching theory of intelligent interactive behaviour that is capable of informing the system-level design of such systems.

This talk addresses these issues and argues that we need to go far beyond our current capabilities and understanding towards a more integrated perspective.

We need to move from developing devices that simply talk and listen to evolving intelligent communicative machines that are capable of truly understanding human behaviour.