Human Interactivity and Language Lab

IEEE Robotics & Automation Society

Speakers and tutors | Programme | Practical info | Application | Organizers

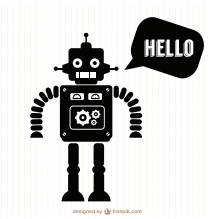
September 18-23, 2023, Poland

4th Summer School on Social Human-Robot Interaction

1

© 2023 The University of Sheffield

Should a Robot Speak?
(*if so, why, when and how*)



Prof. Roger K. Moore

Chair of Spoken Language Processing
Dept. Computer Science, University of Sheffield
(Visiting Prof., Language Sciences, University College London)
(Visiting Prof., Bristol Robotics Lab.)

SHEFFIELD ROBOTICS

University of Sheffield

SPandH

2

© 2023 The University of Sheffield

Talking with Machines

Command and Control Systems

Dictation Systems

Interactive Voice Response (*IVR*) Systems

Voice-Enabled Personal Assistants

Embodied Conversational Agents (*ECAs*)

Autonomous Agents (*robots*)

3

© 2023 The University of Sheffield

Talking with Machines

Command and Control Systems

Dictation Systems

Interactive Voice Response (*IVR*) Systems

Voice-Enabled Personal Assistants

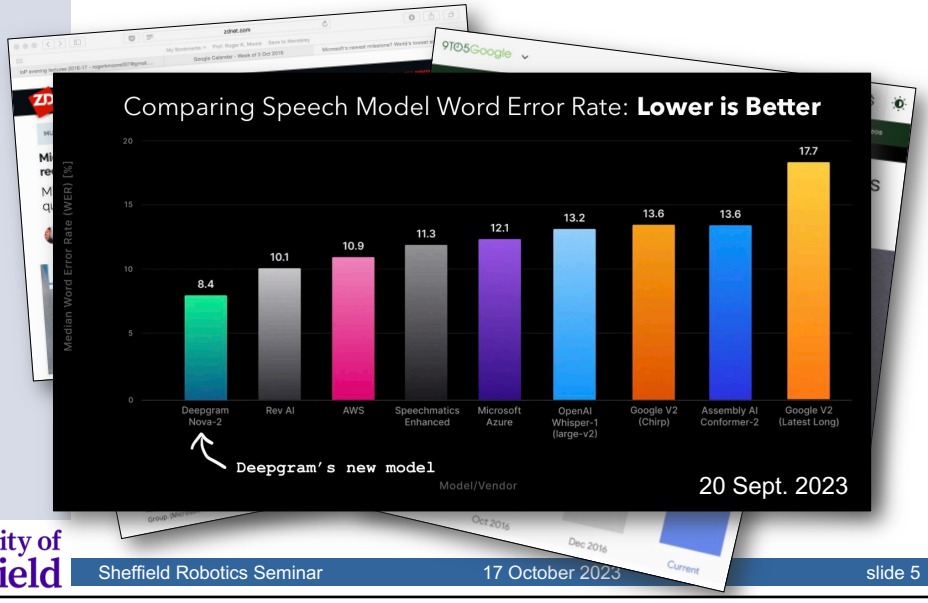
Embodied Conversational Agents (*ECAs*)

Autonomous Agents (*robots*)

4

© 2023 The University of Sheffield

Amazing Progress



5

© 2023 The University of Sheffield

Where Are We Going?

GLOBAL VOICE ASSISTANT MARKET

North America
Largest Market
By Region (2019E)

APAC
Fastest-Growing Market
By Region (2020-2030)

2019 Market Size **$1,723.6** million

2030 Market Size **$26,872.6** million

Market Growth Rate (2020-2030) **29.7 %**

https://www.psmarketresearch.com/market-analysis/voice-assistant-market

6

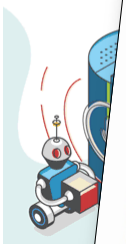
© 2023 The University of Sheffield

Where Are We Going?

GLOBA

North
Largest
By Regi

ng Market
20-2030)



BUSINESS INSIDER UK Tech Finance Politics Strategy All

TECH

More: Innovation robot pepper

The next generation of Siri-like assistants will be robots living in your home

ANTONIO VILLAS-BOAS
JUN. 22, 2015, 6:40 PM

FACEBOOK LINKEDIN TWITTER

Most robots so far have remained on testing platforms or on stages at showcase events to show off a company's technological ability.

The Pepper robot built by Japanese companies Alderbaran and Softbank, on the other hand, can be bought online and used at home right now. Except you may have to wait a long time as the first batch of 1000 Peppers was sold out in about a minute, and it's only available in Japan.

Pepper, the personal digital assistant robot.

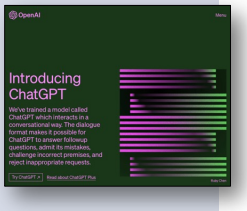
Market Growth Rate (2020-2030) 29.7 %

https://www.psmarket...
-analysis/voice-assistant-market

7

© 2023 The University of Sheffield

Recent Trends



OpenAI Research API ChatGPT Safety Company

25 Sept 2023

ChatGPT can now see, hear, and speak

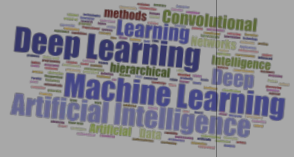
We are beginning to roll out new voice and image capabilities in ChatGPT. They offer a new, more intuitive type of interface by allowing you to have a voice conversation or show ChatGPT what you're talking about.

8

© 2023 The University of Sheffield

The 'State-of-the-Art'

- There is steady year-on-year progress
- Improvements come from:
  - increase in available computer power
  - latest machine-learning paradigms
  - huge real-world training corpora
  - open tools / benchmark testing / challenges
- 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*)
  - inefficient (*trained on more data than a person hears in a lifetime*)
  - biased (*trained on un-curated data*)
  - untrustworthy (*e.g. hallucinates misinformation*)
  - ecologically damaging (*model training has a significant carbon footprint*)
- It is easy to underestimate the richness and complexity of spoken language interaction
- It is not 'natural' to talk to a machine!
- The availability of open tools and data is de-skilling

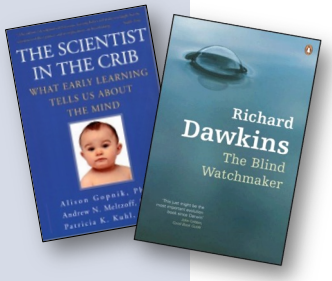


9

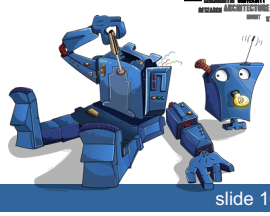
© 2023 The University of Sheffield

Speech is not just another Modality

"Spoken language is the most sophisticated behaviour of the most complex organism in the known universe!"



Simply interfacing state-of-the-art speech technology with a state-of-the-art robot does not lead to effective human-robot interaction (*talking and listening is not enough*)



10

© 2023 The University of Sheffield

Speech is not just Audible Text

communicative

ostensive

situated

embodied

pragmatic

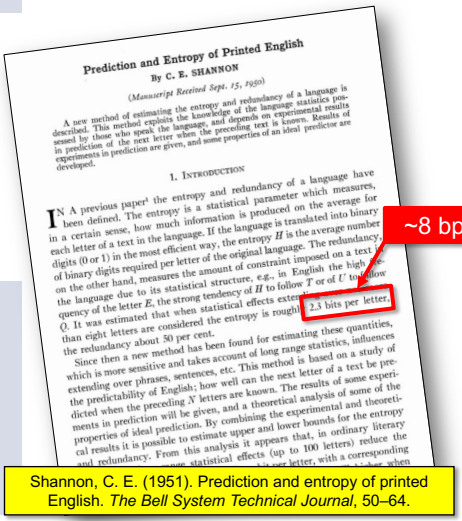
grounded

Speech is …

- variable
- ambiguous
- effortful
- contrastive
- prosodic
- adaptive
- context-dependent
- meaningful
- referential
- indexical
- rhetorical
- paralinguistic
- personalised
- affective
- multimodal
- contaminated

© 2023 The University of Sheffield

Spoken Language > Text



Prediction and Entropy of Printed English

By C. E. SHANNON

(Manuscript Received Sept. 15, 1950)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

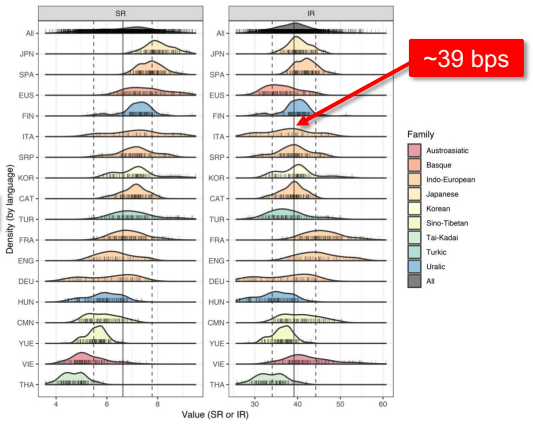
1. Introduction

IN A previous paper[1] the entropy and redundancy of a language have been defined. The entropy is a statistical parameter which measures, in a certain sense, how much information is produced on the average for each letter of a text in the language. If the language is translated into binary digits (0 or 1) in the most efficient way, the entropy H is the average number of binary digits required per letter of the original language. The redundancy, on the other hand, measures the amount of constraint imposed on a text in the language due to its statistical structure, e.g., in English the high frequency of the letter E, the strong tendency of H to follow T or of U to follow Q. It was estimated that when statistical effects extend over not more than eight letters the entropy is roughly 2.3 bits per letter, the redundancy about 50 per cent.

Since then a new method has been found for estimating these quantities, which is more sensitive and takes account of long range statistics, influences extending over phrases, sentences, etc. This method is based on a study of the predictability of English; how well can the next letter of a text be predicted when the preceding N letters are known. The results of some experiments in prediction will be given, and a theoretical analysis of some of the properties of ideal prediction. By combining the experimental and theoretical results it is possible to estimate upper and lower bounds for the entropy and redundancy. From this analysis it appears that, in ordinary literary

~8 bps

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal*, 50–64.



~39 bps

Coupé, C., Oh, Y., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9).

© 2023 The University of Sheffield

The Information Rate of Spoken Language

- Of course, spoken language is *not* a fixed code with a constant information rate
- What people say is conditioned on critical *causal* factors such as ...
  - their situated and embodied circumstances: '*pragmatics*'
  - the temporal evolution of events: '*synchronics*'
  - the level of effort that they are prepared to devote to their behaviour: '*energetics*'
- In other words, the information rate in spoken language varies as a function of the pragmatic, synchronic and energetic context

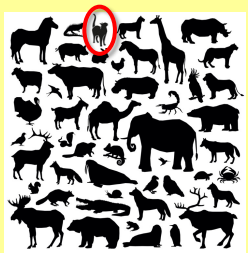
Moore, R. K.. (2023). Pragmatics, synchronics and energetics in spoken language – an information theoretic perspective. In *Workshop on Limits and Benefits of Information-Theoretic Perspectives in Spoken Communication* (LBIT), Dublin.

13

© 2023 The University of Sheffield

Pragmatics: *situational context*

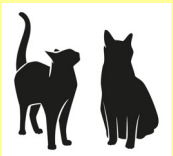
Entropy = 5.58 bits



"*Ooh, look at that cat*"

Communicated = 5.58 bits

Entropy = 1 bit



"*Ooh, look at that cat … the one on the left*"

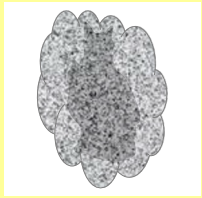
Communicated = 1 bit

Entropy = 1 bit

"*Ooh, look at that cat … the one in front and slightly to the left*"

Communicated = ~80 bits

Entropy = 1 bit



"*Ooh, look at that … I think it's a cat, or maybe two cats with one sitting in front of the other as I can see what looks like three ears with one head lower than the other and slightly to the left?*"

Communicated = ~440 bits

Silhouette images from Freepik

14

© 2023 The University of Sheffield

Energetics: *motivational context*

Transmitted Information ∝ Effort

"*Why do I have to shout?*"

"Where's the newspaper?"

"I don't know"

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31, 373-405.

"*Huh? Speak more clearly!*"

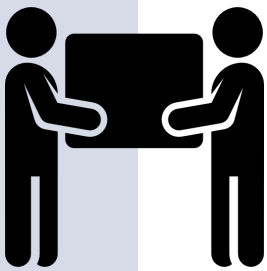
Silhouette images from Freepik

15

© 2023 The University of Sheffield

Synchronics: *temporal context*

"*To me*" … "*To you*"



"*Three, two, one … GO!*"

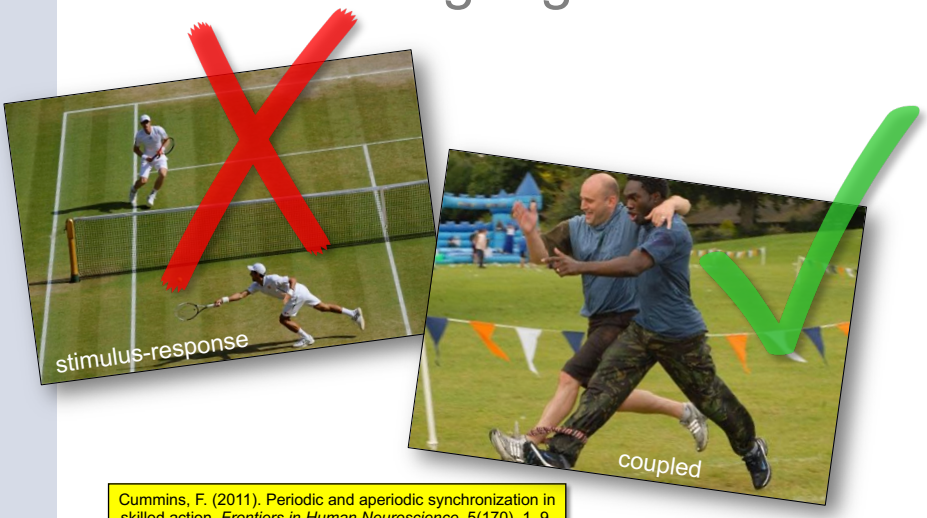
"*Look at that …*"

Silhouette images designed by kjpargeter / Freepik

16

© 2023 The University of Sheffield

What is Language Like?

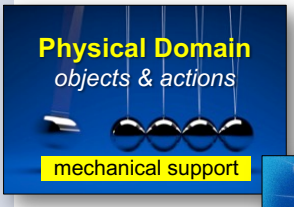


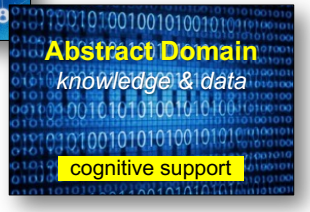
Cummins, F. (2011). Periodic and aperiodic synchronization in skilled action. *Frontiers in Human Neuroscience*, 5(170), 1–9.

17

© 2023 The University of Sheffield

Why Use Spoken Language?





hands-free manipulation

eyes-free observation

collaborative coordination

the management of joint attention

multi-modal interaction

the expression of empathy

the expression of companionship

verbal + non-verbal signalling

the expression of emotion

sense making

access to information

high-rate communications

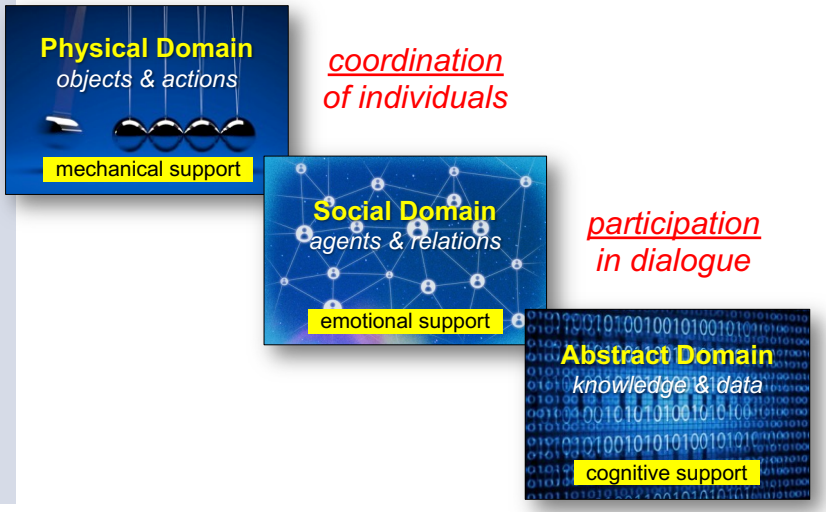
problem solving

Moore, R. K. (2019). Talking with Robots: Opportunities and Challenges. *International Conference Language Technologies for All (LT4All).*

18

© 2023 The University of Sheffield

When to Use Spoken Language?

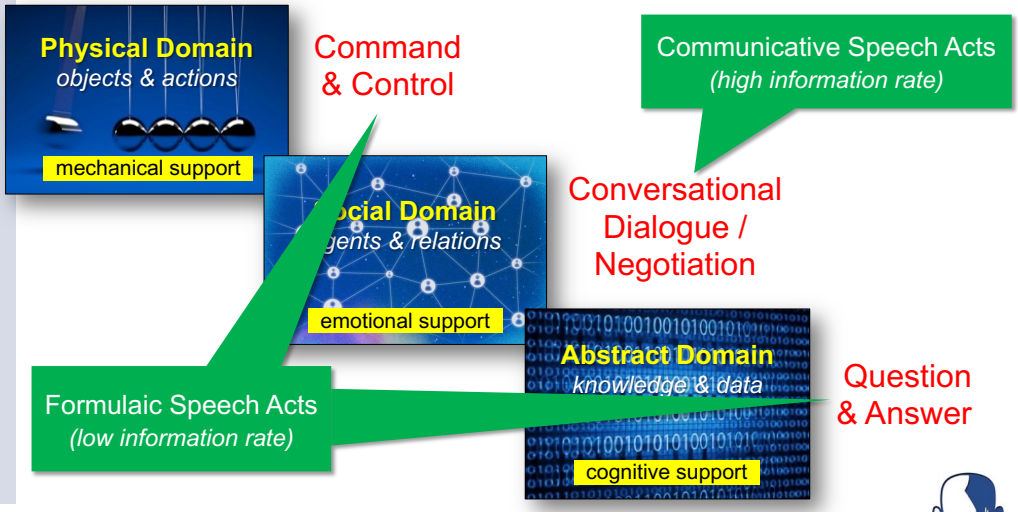


Moore, R. K. (2019). Talking with Robots: Opportunities and Challenges. *International Conference Language Technologies for All (LT4All).*

19

© 2023 The University of Sheffield

How to Use Spoken Language?



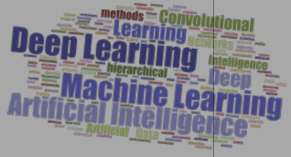
Moore, R. K. (2019). Talking with Robots: Opportunities and Challenges. *International Conference Language Technologies for All (LT4All).*

20

© 2023 The University of Sheffield

The 'State-of-the-Art'

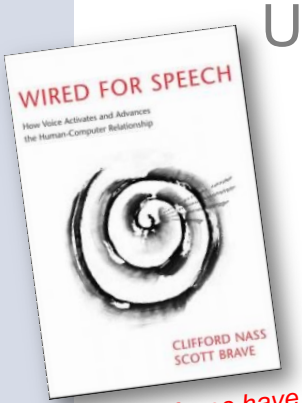
- There is steady year-on-year progress
- Improvements come from:
  - increase in available computer power
  - latest machine-learning paradigms
  - huge real-world training corpora
  - open tools / benchmark testing / challenges
- Progress has not come about as a result of deep insights into human spoken language
- Spoken language technology is …
  - fragile (*in 'real' conditions*)
  - expensive (*to port to new applications / languages*)
  - inefficient (*trained on more data than a person hears in a lifetime*)
  - biased (*trained on un-curated data*)
  - untrustworthy (*e.g. hallucinates misinformation*)
  - ecologically damaging (*model training has a significant carbon footprint*)
- It is easy to underestimate the richness and complexity of spoken language interaction
- It is not 'natural' to talk to a machine!
- The availability of open tools and data is de-skilling



21

© 2023 The University of Sheffield

Usability Issues



"Voice interfaces have become notorious for fostering frustration and failure."

noise
accents
understanding
privacy
embarrassment
alternative GUIs
task familiarity
limited functionality

How many 'real' users?

Moore, R. K., Li, H., & Liao, S.-H. (2016). Progress and prospects for spoken language technology: what ordinary people think. *INTERSPEECH* (pp. 3007–3011). San Francisco, CA.

22

© 2023 The University of Sheffield

Alexa is not a Robot!

(*and a robot may not be a personal assistant*)

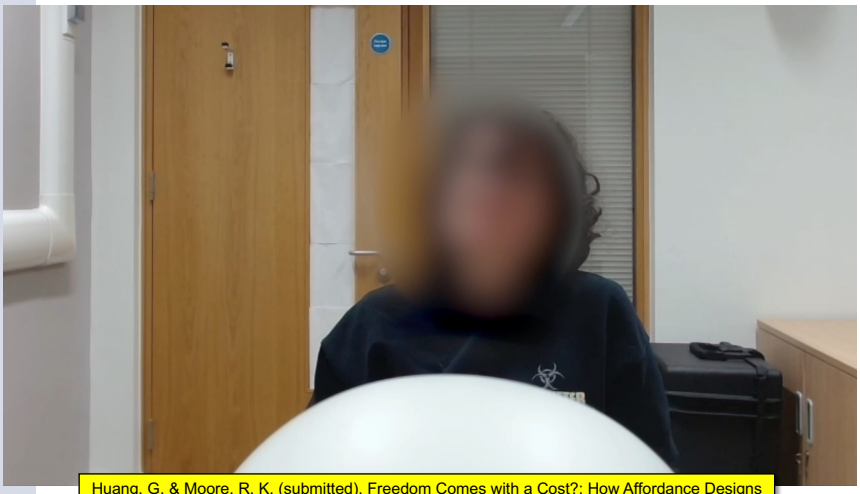
"A robot is an actuated mechanism programmable in two or more axes with a degree of autonomy, moving within its environment, to perform intended tasks."

http://www.leorobotics.nl/definition-robots-and-robotics

23

© 2023 The University of Sheffield

Talking to a Robot

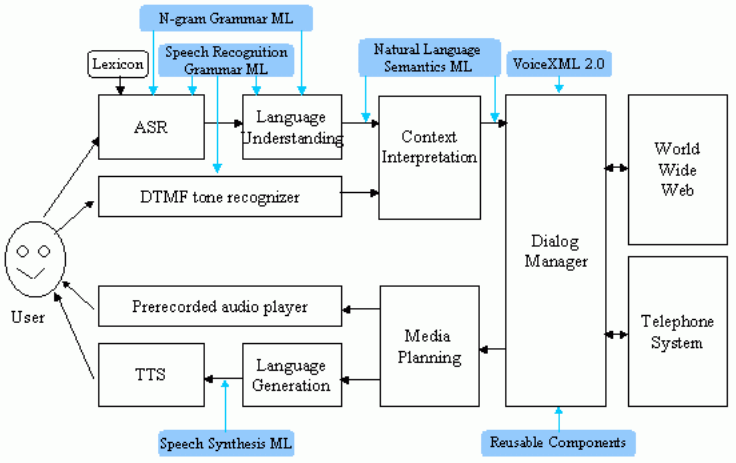


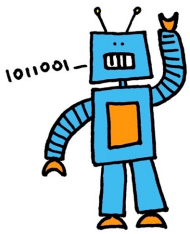
Huang, G. & Moore, R. K. (submitted). Freedom Comes with a Cost?: How Affordance Designs Affect Users' Experience with a Conversational Social Robot. *Frontiers in Robotics and AI*.

24

© 2023 The University of Sheffield

‘Standard’ Speech Interface



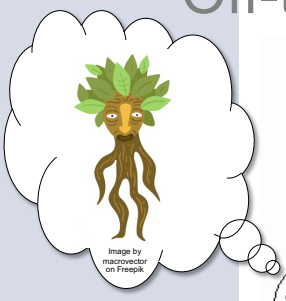


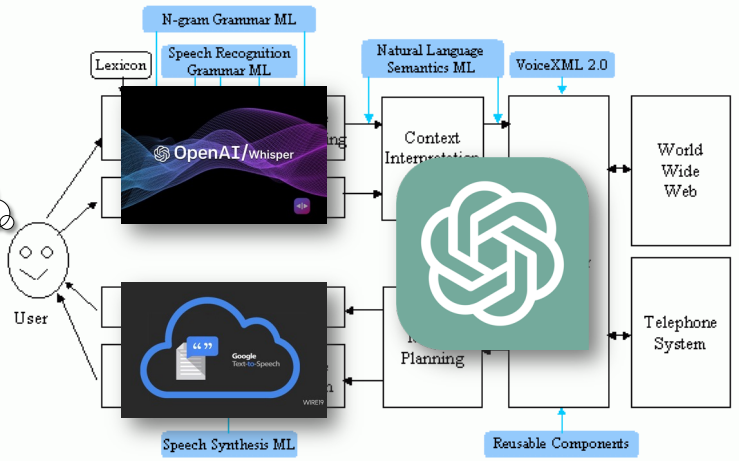
Introduction and Overview of W3C Speech Interface Framework, http://www.w3.org/TR/voice-intro/

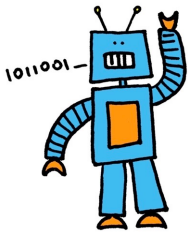
25

© 2023 The University of Sheffield

‘Off-the-Shelf’ Speech Components



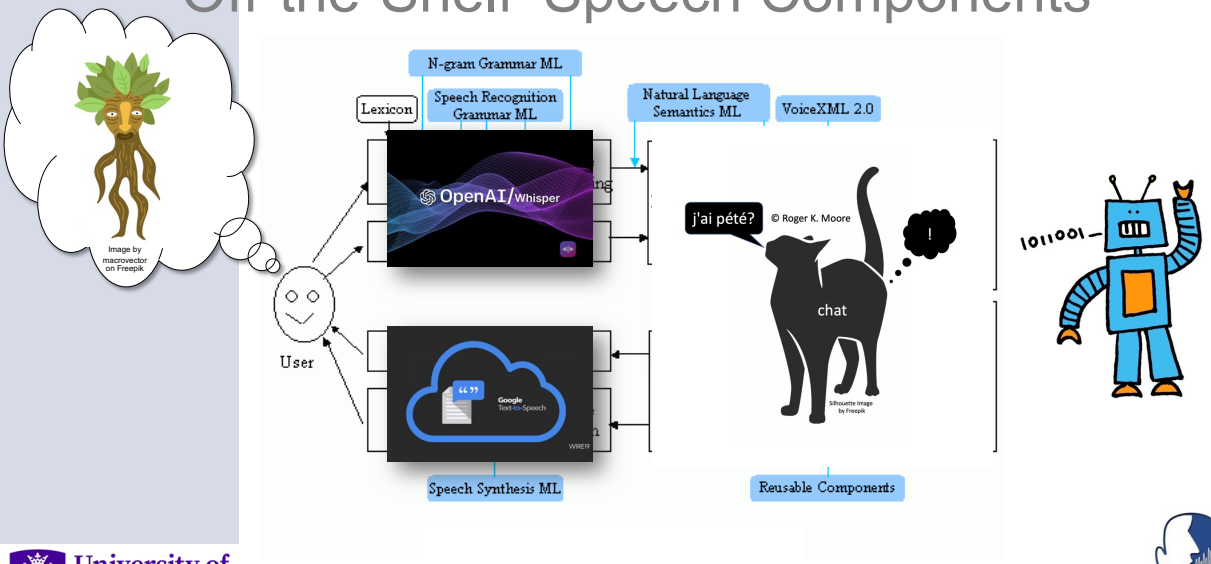




26

© 2023 The University of Sheffield

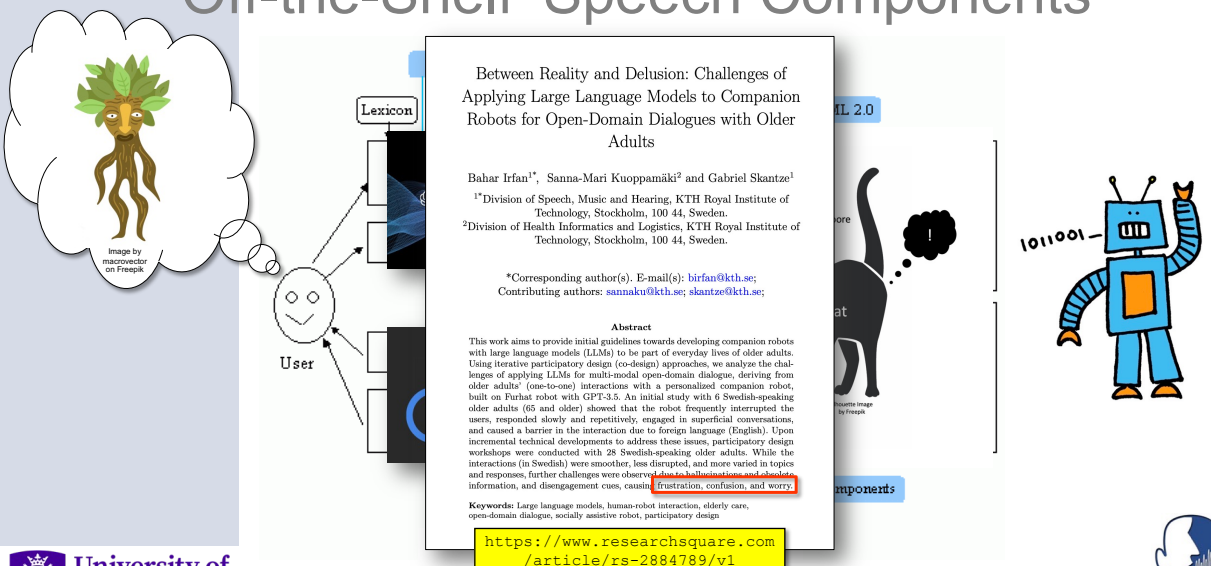
‘Off-the-Shelf’ Speech Components



27

© 2023 The University of Sheffield

‘Off-the-Shelf’ Speech Components



Between Reality and Delusion: Challenges of Applying Large Language Models to Companion Robots for Open-Domain Dialogues with Older Adults

Bahar Irfan[1*], Sanna-Mari Kuoppamäki[2] and Gabriel Skantze[1]
[1*]Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, 100 44, Sweden.
[2]Division of Health Informatics and Logistics, KTH Royal Institute of Technology, Stockholm, 100 44, Sweden.
*Corresponding author(s). E-mail(s): birfan@kth.se;
Contributing authors: sannaku@kth.se; skantze@kth.se;

Abstract
This work aims to provide initial guidelines towards developing companion robots with large language models (LLMs) to be part of everyday lives of older adults. Using iterative participatory design (co-design) approaches, we analyze the challenges of applying LLMs for multi-modal open-domain dialogue, deriving from older adults' (one-to-one) interactions with a personalized companion robot, built on Furhat robot with GPT-3.5. An initial study with 6 Swedish-speaking older adults (65 and older) showed that the robot frequently interrupted the users, responded slowly and repetitively, engaged in superficial conversations, and caused a barrier in the interaction due to foreign language (English). Upon incremental technical developments to address these issues, participatory design workshops were conducted with 28 Swedish-speaking older adults. While the interactions (in Swedish) were smoother, less disrupted, and more varied in topics and responses, further challenges were observed due to hallucinations and obsolete information, and disengagement cues, causing frustration, confusion, and worry.

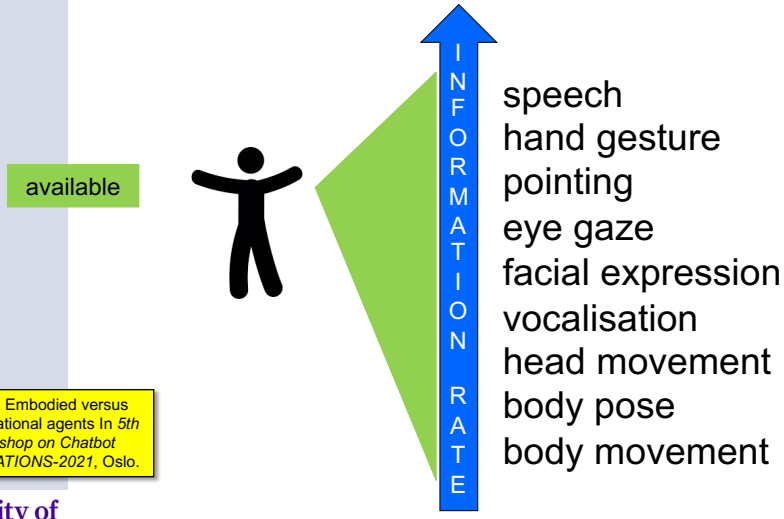
Keywords: Large language models, human-robot interaction, elderly care, open-domain dialogue, socially assistive robot, participatory design

https://www.researchsquare.com/article/rs-2884789/v1

28

© 2023 The University of Sheffield

Multi-Modal Communication

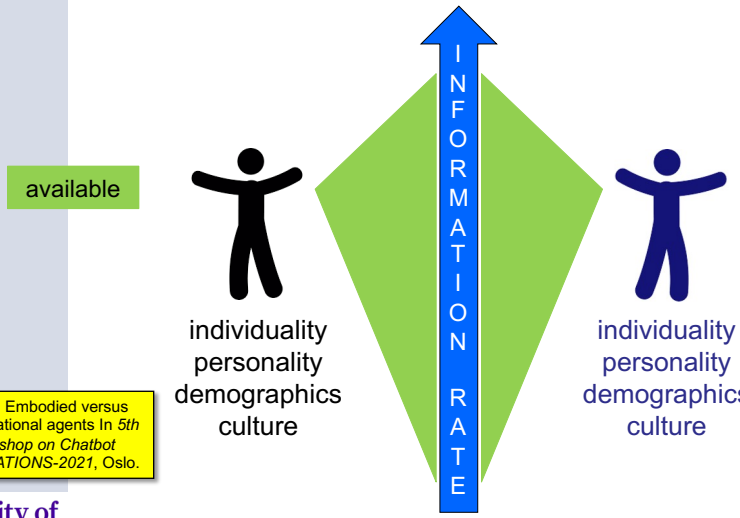


Moore, R. K. (2021). Embodied versus disembodied conversational agents In *5th International Workshop on Chatbot research: CONVERSATIONS-2021*, Oslo.

29

© 2023 The University of Sheffield

Multi-Modal Human-Human Interaction

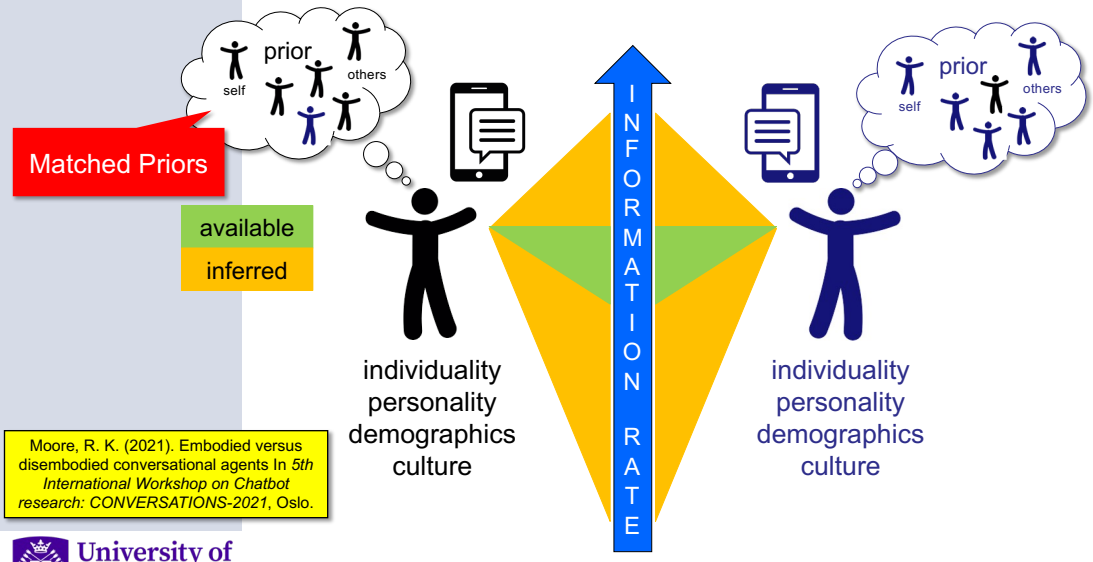


Moore, R. K. (2021). Embodied versus disembodied conversational agents In *5th International Workshop on Chatbot research: CONVERSATIONS-2021*, Oslo.

30

© 2023 The University of Sheffield

Mono-Modal Human-Human Interaction

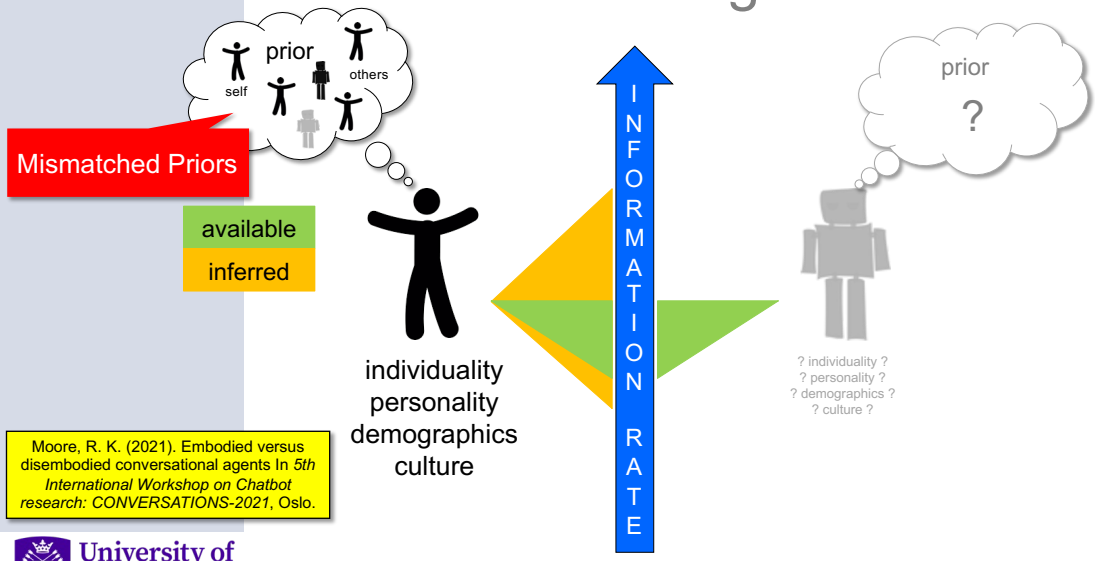


Moore, R. K. (2021). Embodied versus disembodied conversational agents In *5th International Workshop on Chatbot research: CONVERSATIONS-2021*, Oslo.

31

© 2023 The University of Sheffield

Mono-Modal Human-Agent Interaction

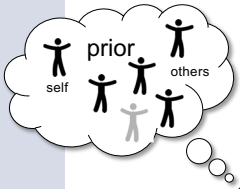


Moore, R. K. (2021). Embodied versus disembodied conversational agents In *5th International Workshop on Chatbot research: CONVERSATIONS-2021*, Oslo.

32

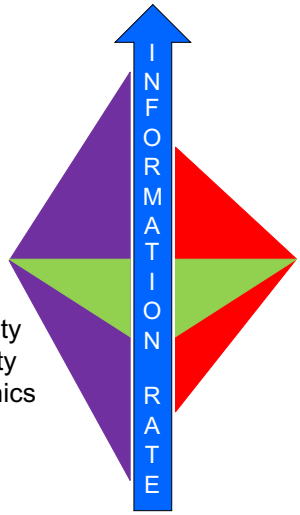
© 2023 The University of Sheffield

Multi-Modal Human-Agent Interaction



available
deception
confusion

individuality
personality
demographics
culture



? individuality ?
? personality ?
? demographics ?
? culture ?

As well as using human-like bodies, faces and voices, it is easy to fake …

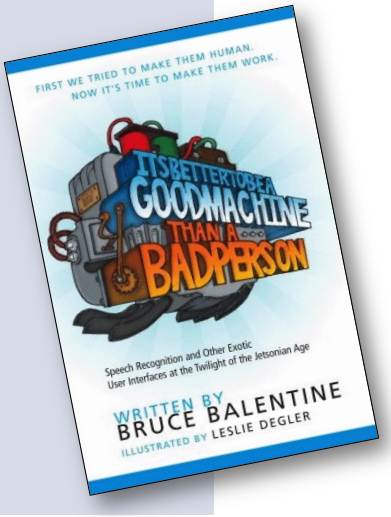
- emotion
- empathy
- understanding
- sympathy
- opinions
- likes/dislikes
- etc. etc.

Moore, R. K. (2021). Embodied versus disembodied conversational agents In *5th International Workshop on Chatbot research: CONVERSATIONS-2021*, Oslo.

33

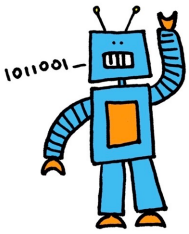
© 2023 The University of Sheffield

Mismatched Interlocutors



- Spoken language interaction between human beings is founded on shared experiences, representations and priors (*i.e. mutual situatedness/embodiment*)
- So, is there a fundamental limit to the language-based interaction that can take place between mismatched interlocutors?

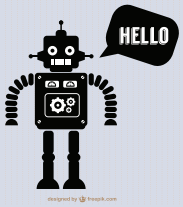
Moore, R. K. (2016). Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In K. Jokinen & G. Wilcock (Eds.), *Dialogues with Social Robots – Enablements, Analyses, and Evaluation*. Springer Lecture Notes in Electrical Engineering (LNEE).



34

© 2023 The University of Sheffield

Talking to a Machine

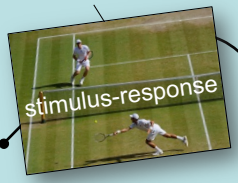


Philips, M. (2006). Applications of spoken language technology and systems. In M. Gilbert & H. Ney (Eds.), *IEEE/ACL Workshop on Spoken Language Technology* (SLT)

Like a Human

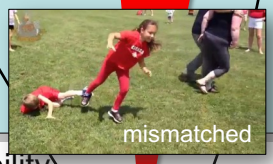
Usability

Add NL/Dialog



'Habitability Gap'

Structured Dialog

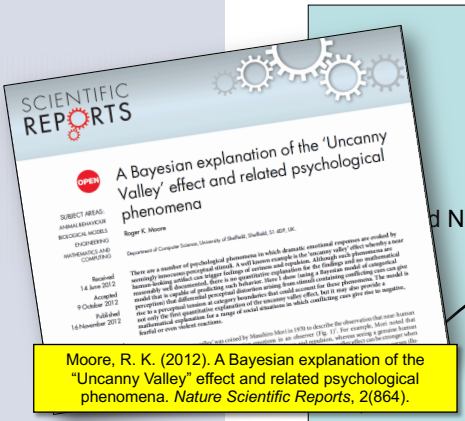


Flexibility

35

© 2023 The University of Sheffield

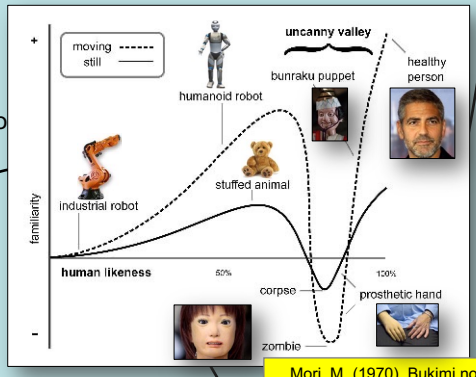
Talking to a Machine



Moore, R. K. (2012). A Bayesian explanation of the "Uncanny Valley" effect and related psychological phenomena. *Nature Scientific Reports*, 2(864).

Like a Human

d NL/Dialo



Structured Dialog

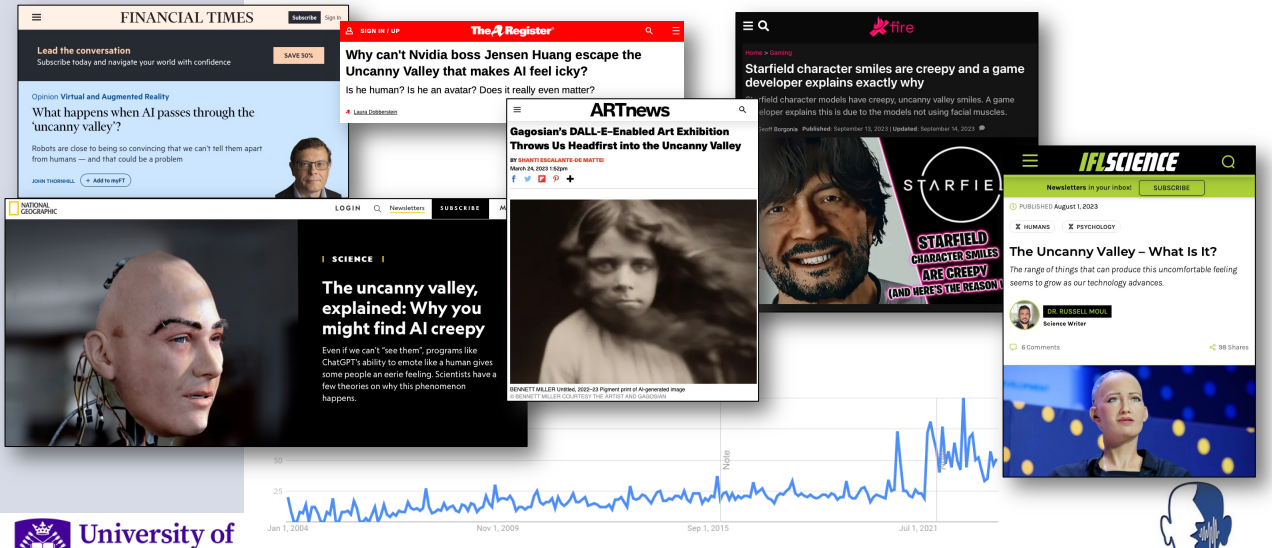
Mori, M. (1970). Bukimi no tani (the uncanny valley). *Energy*, 7, 33-35.

Flexibility

36

© 2023 The University of Sheffield

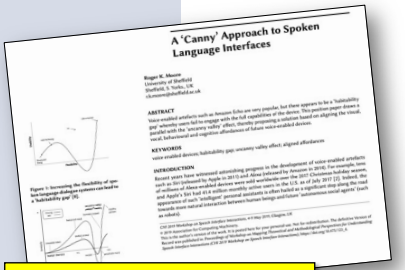
The Uncanny Valley



37

© 2023 The University of Sheffield

A 'Canny' Approach



Moore, R. K. (2019). A 'Canny' Approach to Spoken Language Interfaces. *CHI-19 Workshop on Mapping Theoretical and Methodological Perspectives for Understanding Speech Interface Interactions*, Glasgow.

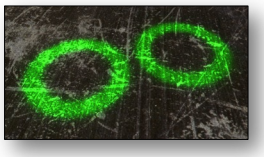
- Minimise conflicting cues by aligning an agent's 'affordances':
  - visual (*i.e. what it looks like*)
  - vocal (*i.e. what it sounds like*)
  - behavioural (*i.e. how it behaves*)
  - cognitive (*i.e. what it appears to know*)
  - communicative (*i.e. how it interacts*)
- This means that an agent's capabilities will be capped (*but in a meaningful way*)
- It also implies that a voice-enabled agent should have an 'appropriate' voice (*and a chatbot should use an appropriate font*!)

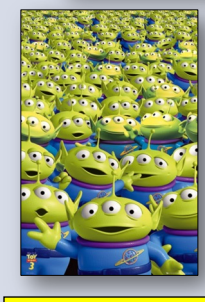
uncanniness ↓ ➤ affinity ↑ ➤ rapport ↑

38

© 2023 The University of Sheffield

‘Appropriate’ Voices





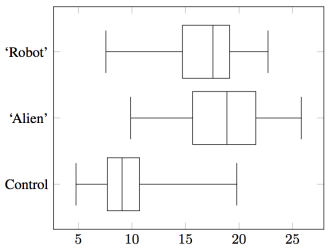
Wilson, S., & Moore, R. K. (2017). Robot, alien and cartoon voices: implications for speech-enabled systems. *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots* (VIHAR-2017), 40–44.

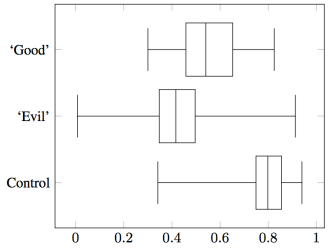
39

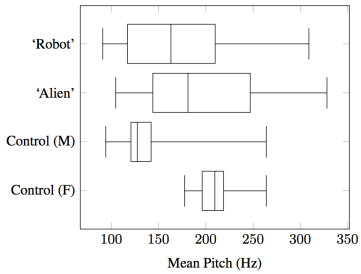
© 2023 The University of Sheffield

Designer Voices

Technique	Method
Time reversal	*delay line*
Speed change	*delay line*
Tremolo	*modulated amplitude*
Vibrato	*modulated pitch*
Ring modulation	*multiplication of two signals*
Comb filter	*short delayed version added to the original*
Echo	*long delayed version added to the original*
Flanger	*delay-modulated version added to the original*
Chorus	*multiple flangers with different delays*
Phaser	*phase-modulated version added to the original*
Reverberation	*convolution with room acoustic*
Pitch shift	*homomorphic filtering*
Harmony	*pitch-shifted version added to the original*
Filtering	*frequency shaping*
Formant shift	*altered vocal tract length*
Vocoding	*linear prediction analysis-synthesis*





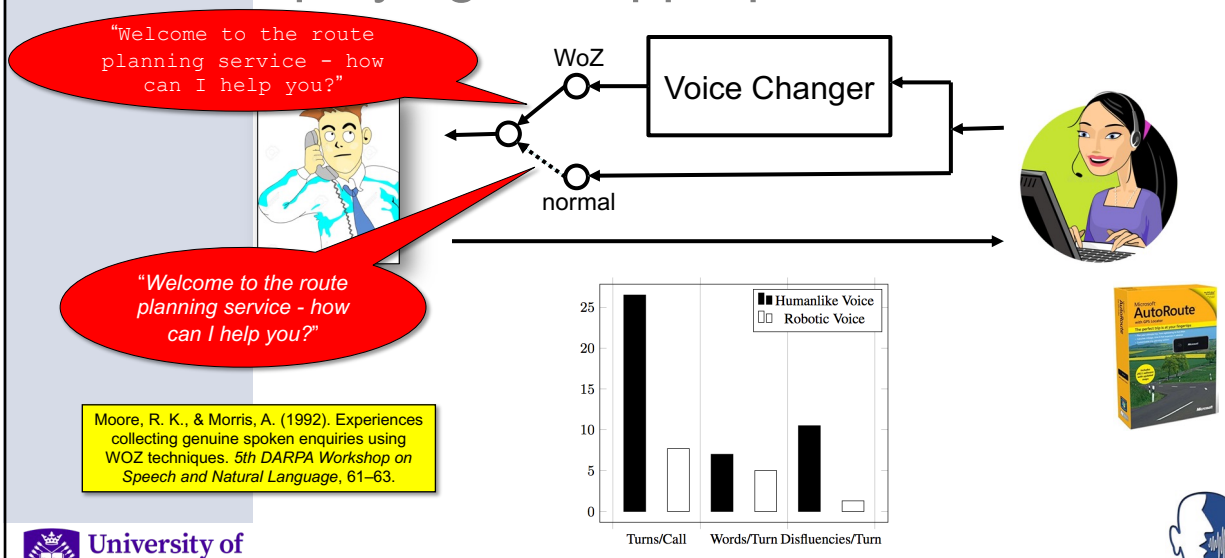


Wilson, S., & Moore, R. K. (2017). Robot, alien and cartoon voices: implications for speech-enabled systems. *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots* (VIHAR-2017), 40–44.

40

© 2023 The University of Sheffield

Deploying an 'Appropriate' Voice

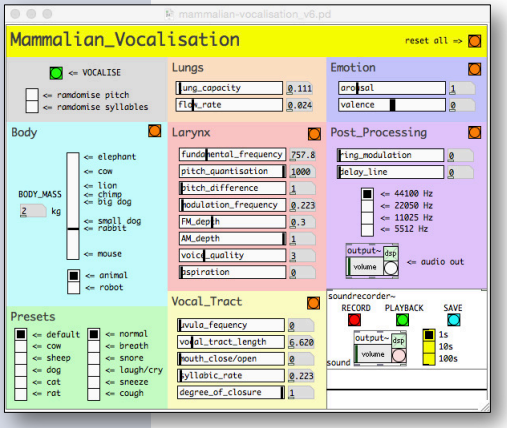


Moore, R. K., & Morris, A. (1992). Experiences collecting genuine spoken enquiries using WOZ techniques. *5th DARPA Workshop on Speech and Natural Language*, 61–63.

41

© 2023 The University of Sheffield

A Voice for the '*MiRo*' Robot



Moore, R. K. (2016). A real-time parametric general-purpose mammalian vocal synthesiser. In *INTERSPEECH* (pp. 2636–2640). San Francisco, CA.

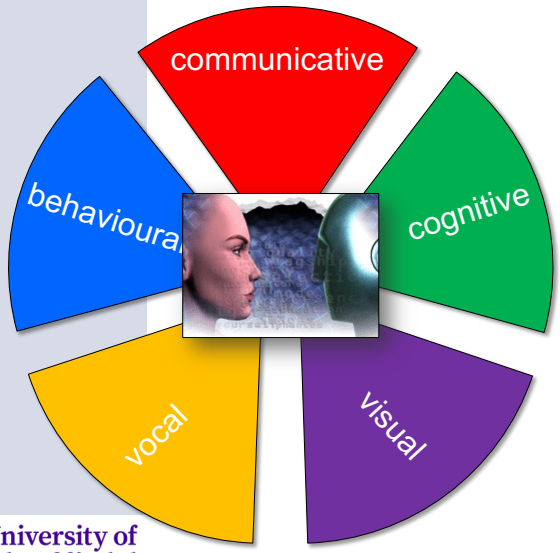
Body mass = 2 kg
Breathing rate = 0.7 Hz
Fundamental frequency = 760 Hz
Vocal tract length = 6.6 cm

Moore, R. K., & Mitchinson, B. (2017). A biomimetic vocalisation system for MiRo. In *Living Machines 2017*. Stanford, CA.

42

© 2023 The University of Sheffield

Bringing It All Together



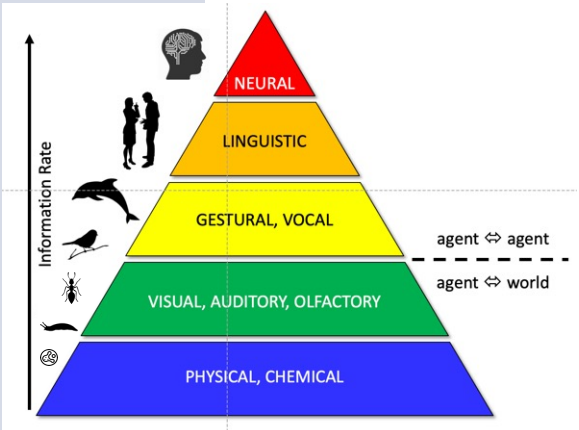
Need to take a holistic approach to …

- designing and implementing interactional affordances
- accommodating user priors

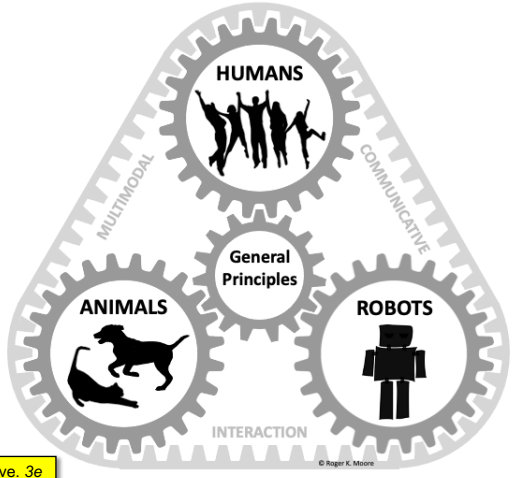
43

© 2023 The University of Sheffield

Bringing It All Together



Moore, R. K. (2021). Où les êtres Sociaux: vers une théorie de l'interaction communicative. *3e Colloque International - Objets Animés, Humains, Animaux : Partenaires de Soins Tendres.*

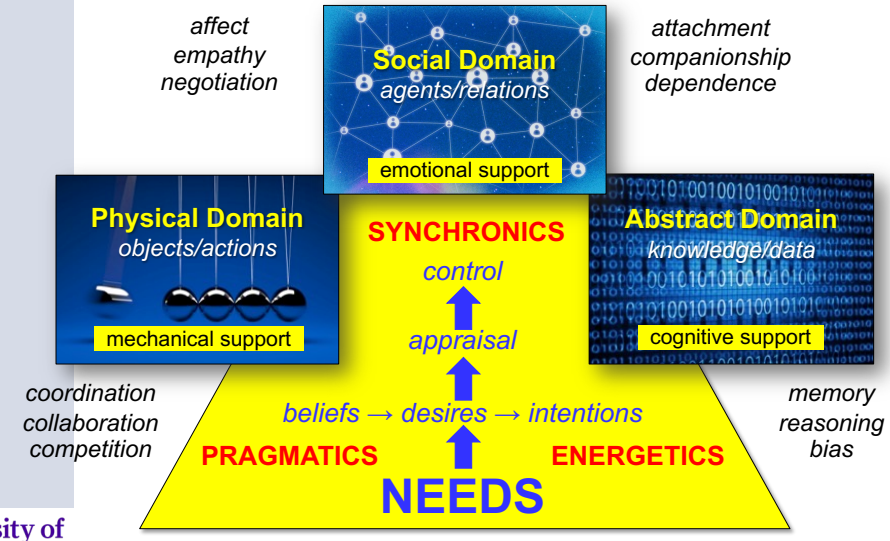


SHARED ENVIRONMENT

44

© 2023 The University of Sheffield

Bringing It All Together

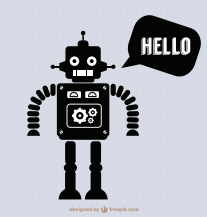


© 2023 The University of Sheffield

Bringing It All Together

© 2023 The University of Sheffield

Summary

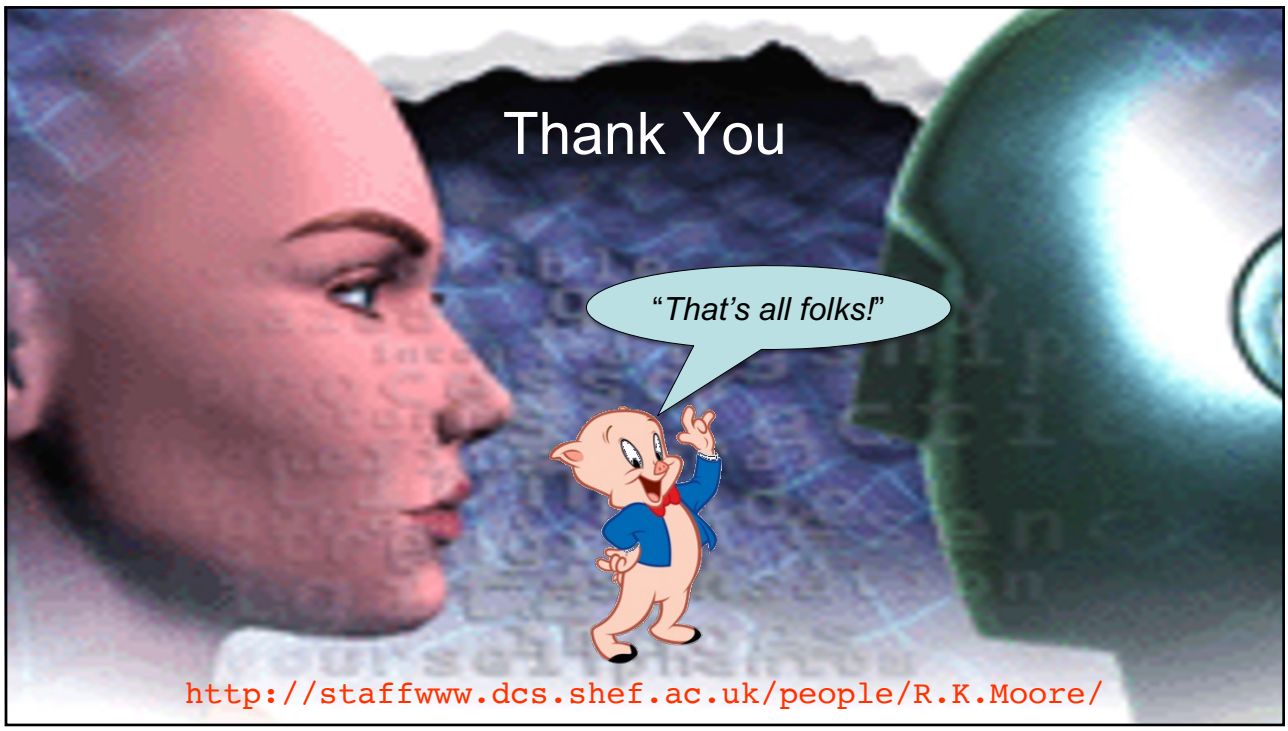


- Next-generation language-based interactive systems need to have a much deeper understanding of user behaviour
- Embodied systems need to facilitate continuous adaptive multimodal incremental closed-loop coupling between users and the physical, social and abstract worlds
- The visual, vocal and behavioural affordances of speech-enabled devices need to reflect their actual abilities
- We need to understand how language might function between mismatched agents
- We need to be able to measure (*and thus optimise and predict*) 'usability' for language-enabled systems
- We must guard against inappropriate use-cases (*e.g. infantalising users*)

© 2023 The University of Sheffield

Where to Find Out More

https://youtu.be/01ml3rpPZYg



49

© 2023 The University of Sheffield

Abstract

- It is often taken for granted that speech provides a 'natural' means by which users can interact with an autonomous agent such as a robot - especially if it is humanoid in design.
- Indeed, the ready availability of off-the-shelf spoken language tools such as Google's speech-to-text/text-to-speech or OpenAI's Whisper/ChatGPT makes it relatively easy to implement a ready-made solution.
- However, not only does such an approach risk creating a chimera - a potentially confusing and inappropriate compilation of misaligned visual, vocal and behavioural affordances - but it also facilitates a casual misrepresentation/faking of agency (e.g. by having the robot express apparently personal likes, dislikes and opinions using the pronoun "I").
- This talk will address these issues, and provide an insight into why, when and how users might wish to converse with a robot.

- 60 mins = ~30 slides

50