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

SHEFFIELD ROBOTICS

EPSRC Engineering and Physical Sciences Research Council

The University Of Sheffield.

SPandH

© 2016 The University of Sheffield

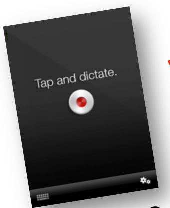
Rich History of Technological Development

Marconi 'SR128' (1982)

Radio Rex (1922)

Apple's "Siri" (2011)

Dragon 'Naturally Speaking' (1997)



Voice dictation on a SmartPhone (2007)

The University Of Sheffield.

SPandH

© 2016 The University of Sheffield
Rich History of Technological Development
Hi there. I'm Cortana.
VoiceTra
Speech To Speech Translator
NICT
Google Now
Ask me anything!
Apple's "Siri" (2011)
Speech-to-Speech Translation
The University Of Sheffield.
UCREL, Lancaster
21st January2016
slide 3
SPandH

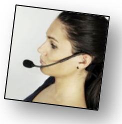
© 2016 The University of Sheffield
Rich History of Technological Development
forbes.com
Forbes
New Posts
Most Popular
Lists
Video
10 Stocks to Buy Now
TECH
Baidu Announces Breakthrough In Speech Recognition, Claiming To Top Google And Apple
+ Comment Now
+ Follow Comments
When artificial-intelligence guru Andrew Ng joined Chinese Internet pioneer Baidu last May as chief scientist, he was a little cagey about what he and his team might work on at a newly opened lab in Sunnyvale, Calif. But he couldn't help revealing better speech recognition as a key area of interest in the age of the smartphone.
Today, Baidu, often called China's Google, unveiled the first results of what the former Google researcher, Stanford professor and Coursera
Forensic Voice Comparison
Voice comparison reports for court LAA rates, Accredited experts
Share
Next Post
The University Of Sheffield.
UCREL, Lancaster
21st January2016
slide 4
SPandH

Past, Present & Future

Command and Control Systems

⇩

Dictation Systems



⇩

Interactive Voice Response (*IVR*) Systems

⇩

Voice-Enabled Personal Assistants

⇩

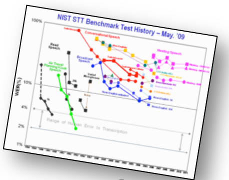
Embodied Conversational Agents (*ECAs*)

⇩

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

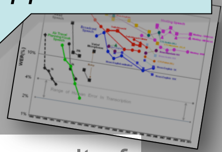


- Progress has *not* come about as a result of deep insights into human spoken language
- Spoken language technology is
  - fragile (*in 'real' conditions*)
  - expensive (*to port to new applications / languages*)
- Performance appears to be reaching an *asymptote that is* well short of human abilities
  - 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 p
  - corpus-driven statistical modelli
  - 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

New opportunities?



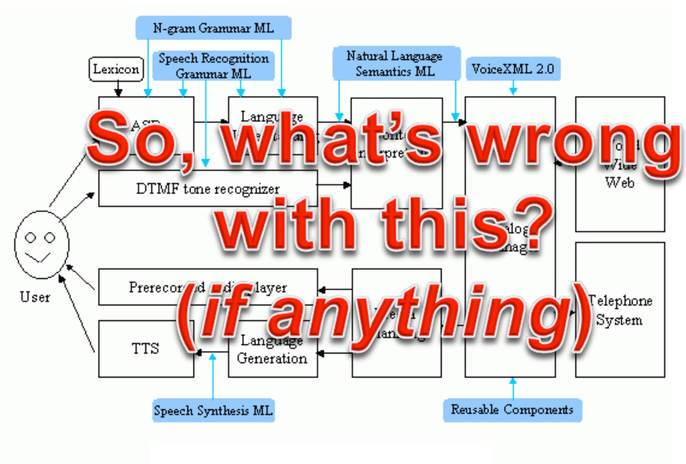
Missed opportunities?

The University Of Sheffield.

SPandH

'Traditional' SLP Architecture

W3C®



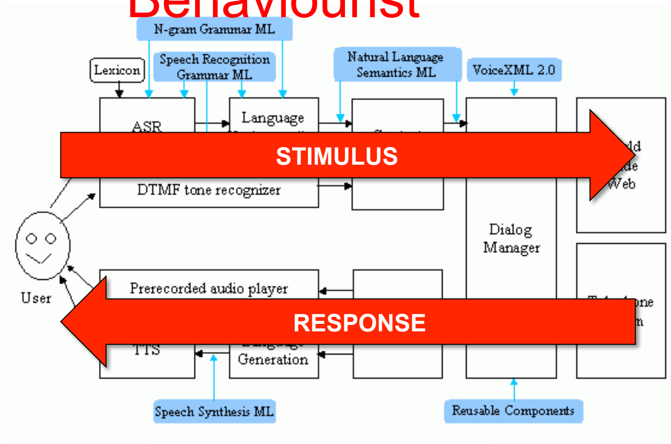
So, what's wrong with this? *(if anything)*

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

The University Of Sheffield.

SPandH

‘Traditional’ ~~SLP~~ Architecture
Behaviourist

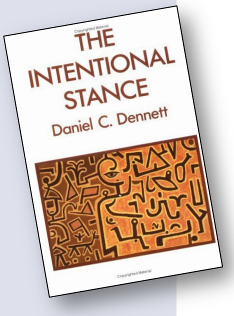


Introduction and Overview of W3C Speech Interface Framework, http://www.w3.org/TR/voice-intro/

The University Of Sheffield.

SPandH

Teleological Behaviour



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

- 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

The University Of Sheffield.

SPandH

Communicating Intentions

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

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

- 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*)

Motivation

This is a ‘compensation’ problem

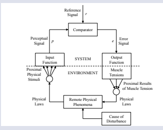
This is an ‘optimisation’ problem

- 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

Perceptual Control Theory



"feedback … is the central and determining factor in all observed behavior"

W. T. Powers (1973). *Behaviour: The Control of Perception*, Aldine, Chicago.

The University Of Sheffield.

SPandH

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*)

The University Of Sheffield.

SPandH

Consequences for SLP

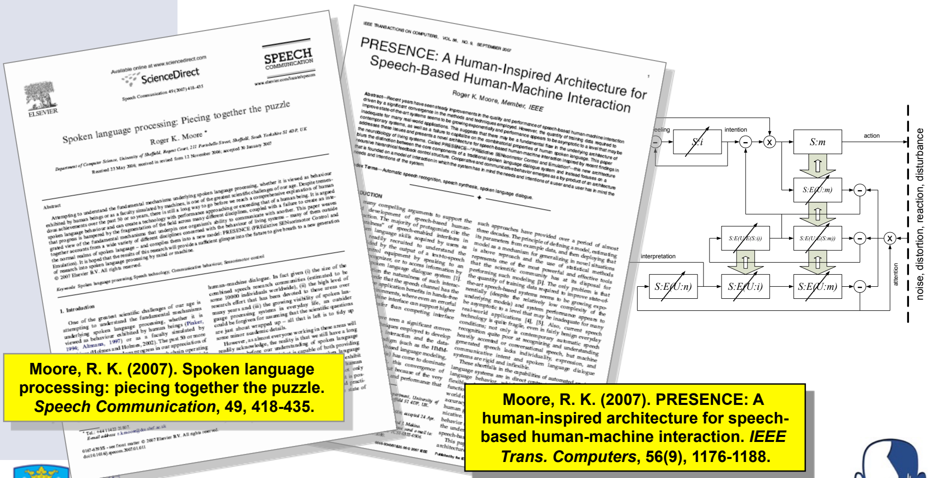
- Need modelling paradigms that are able to accommodate such dependencies
- Emphasises the importance of forward (*generative*) models
- Communicative obstacles are overcome using …
  - sufficient effort
  - feedback
- Communicative effort is related to …
  - the fidelity of the models
  - the depth of the searches

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

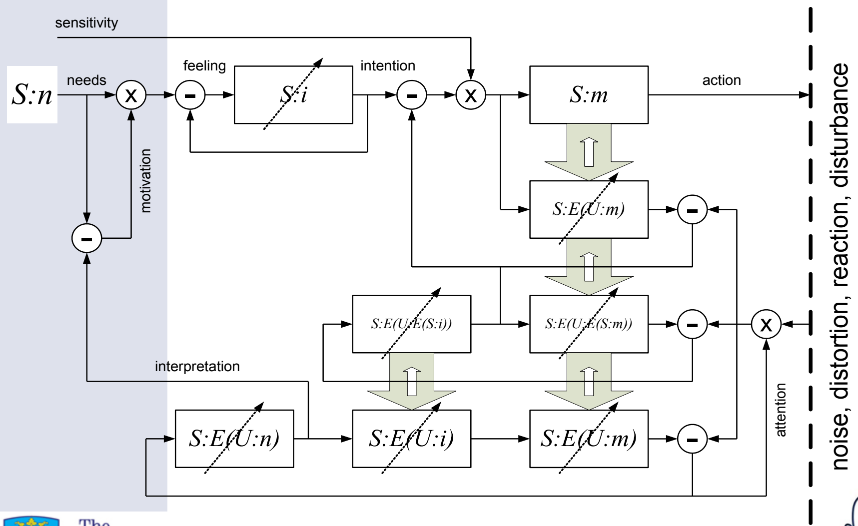
Predictive Sensorimotor Control and Emulation

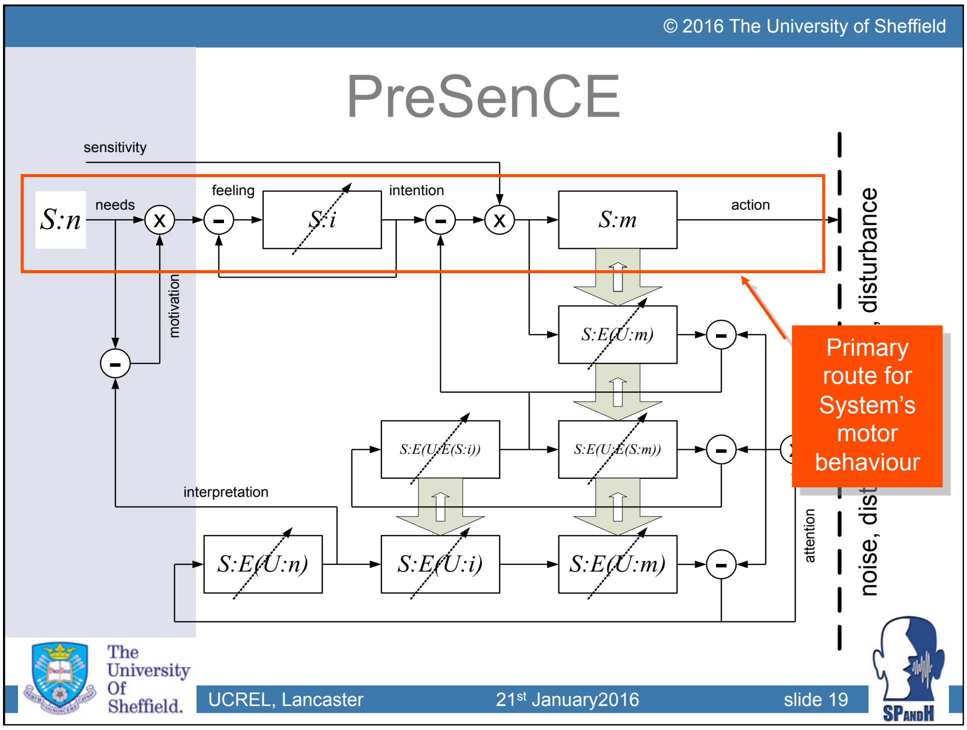


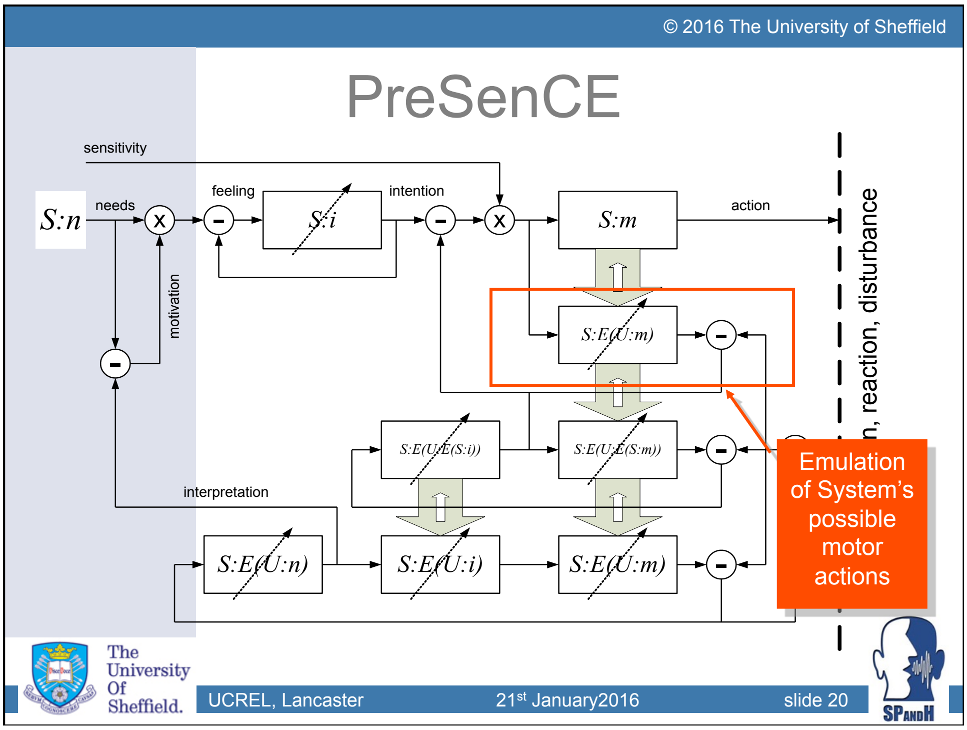
Moore, R. K. (2007). Spoken language processing: piecing together the puzzle. *Speech Communication*, 49, 418-435.

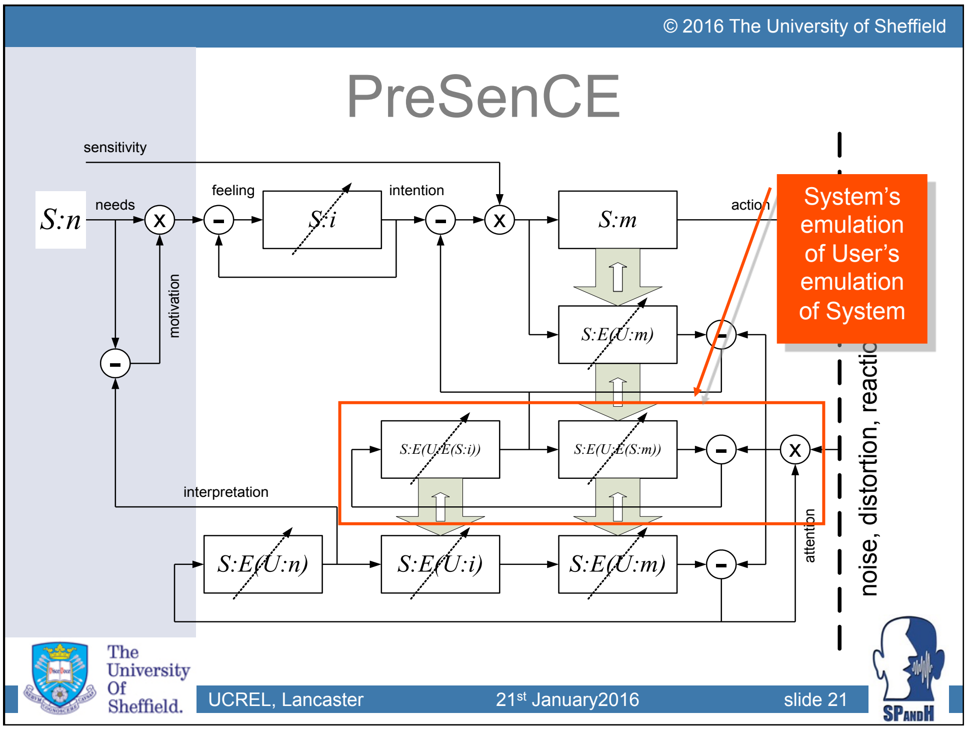
Moore, R. K. (2007). PRESENCE: A human-inspired architecture for speech-based human-machine interaction. *IEEE Trans. Computers*, 56(9), 1176-1188.

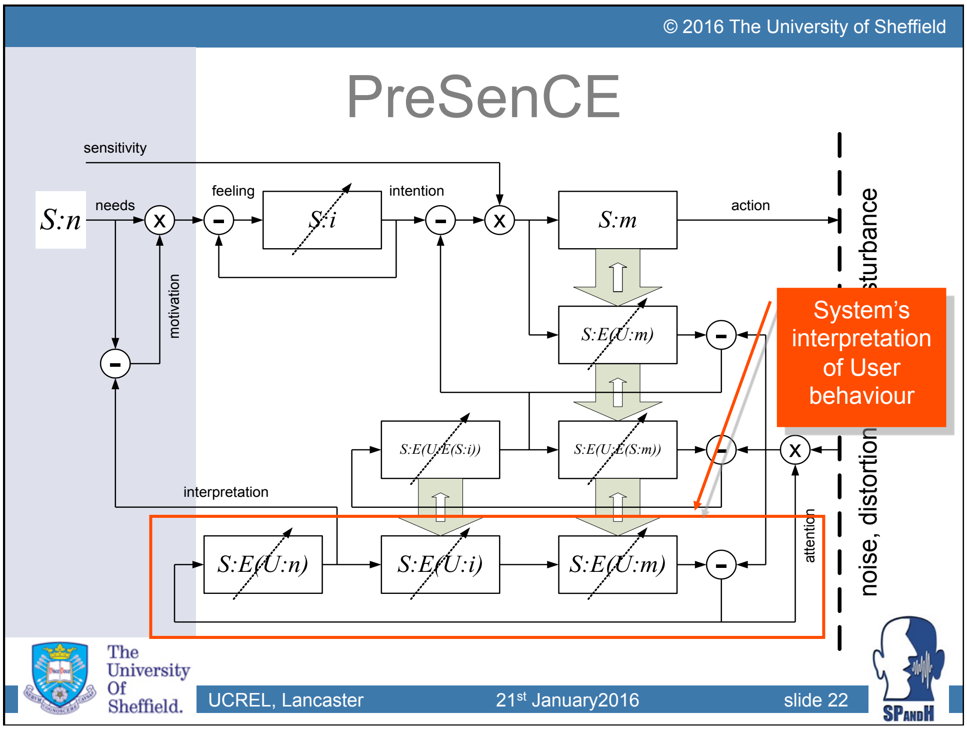
PreSenCE

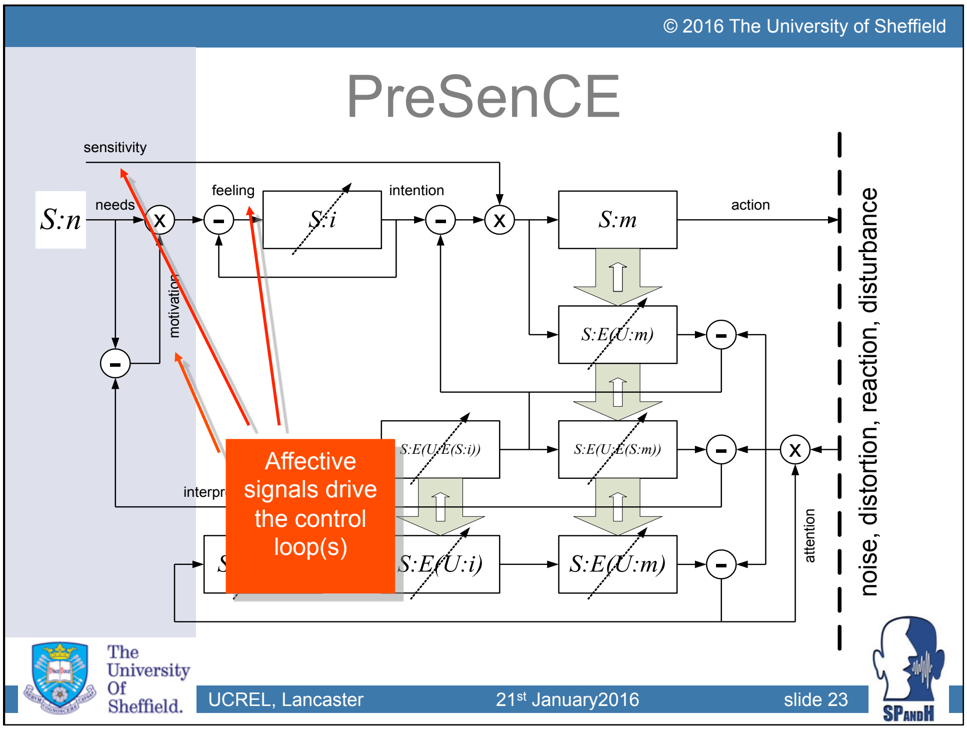


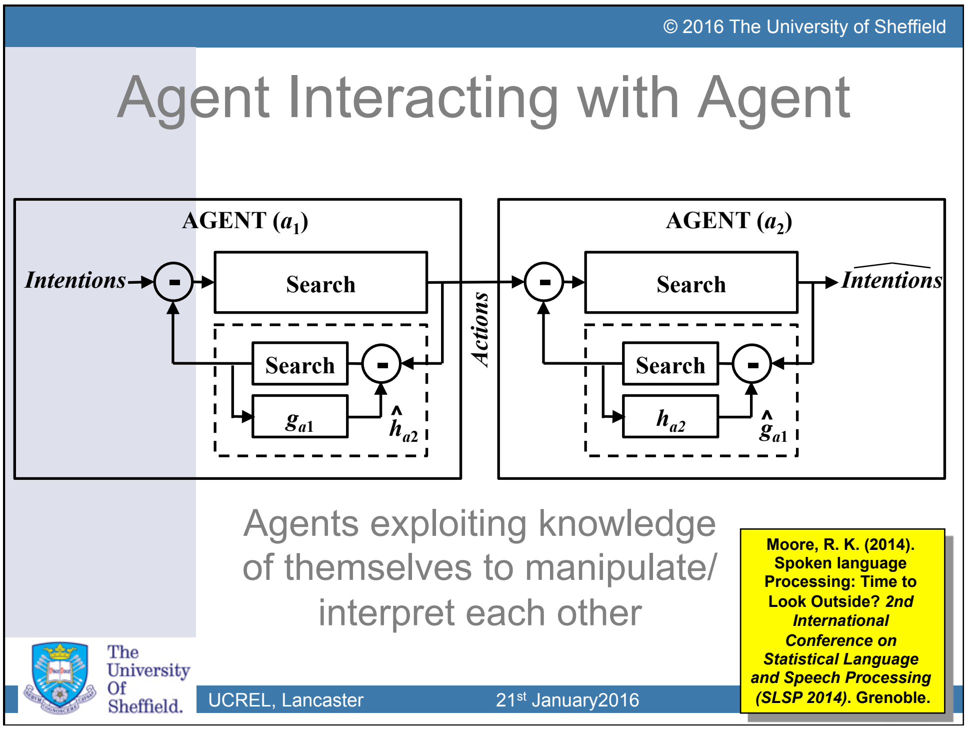

© 2016 The University of Sheffield
PreSenCE
sensitivity
S:n
needs
feeling
S:i
intention
S:m
action
motivation
S:E(U:m)
S:E(U:E(S:i))
S:E(U:E(S:m))
interpretation
S:E(U:n)
S:E(U:i)
S:E(U:m)
attention
noise, dist
, disturbance
Primary route for System's motor behaviour
The University Of Sheffield.
UCREL, Lancaster
21st January2016
slide 19
SPandH


© 2016 The University of Sheffield
PreSenCE
sensitivity
S:n
needs
feeling
S:i
intention
S:m
action
motivation
S:E(U:m)
S:E(U:E(S:i))
S:E(U:E(S:m))
interpretation
S:E(U:n)
S:E(U:i)
S:E(U:m)
n, reaction, disturbance
Emulation of System's possible motor actions
The University Of Sheffield.
UCREL, Lancaster
21st January2016
slide 20
SPandH


© 2016 The University of Sheffield
PreSenCE
sensitivity
S:n
needs
feeling
S:i
intention
S:m
action
motivation
S:E(U:m)
S:E(U:E(S:i))
S:E(U:E(S:m))
interpretation
S:E(U:n)
S:E(U:i)
S:E(U:m)
attention
noise, distortion, reactio
System's emulation of User's emulation of System
The University Of Sheffield.
UCREL, Lancaster
21st January2016
slide 21
SPandH


© 2016 The University of Sheffield
PreSenCE
sensitivity
S:n
needs
feeling
S:i
intention
S:m
action
motivation
S:E(U:m)
S:E(U:E(S:i))
S:E(U:E(S:m))
interpretation
S:E(U:n)
S:E(U:i)
S:E(U:m)
attention
noise, distortion
sturbance
System's interpretation of User behaviour
The University Of Sheffield.
UCREL, Lancaster
21st January2016
slide 22
SPandH


© 2016 The University of Sheffield
PreSenCE
sensitivity
S:n
needs
feeling
S:i
intention
S:m
action
motivation
S:E(U:m)
S:E(U:E(S:i))
S:E(U:E(S:m))
interpr
Affective signals drive the control loop(s)
S:E(U:i)
S:E(U:m)
attention
noise, distortion, reaction, disturbance
The University Of Sheffield.
UCREL, Lancaster
21st January2016
slide 23
SPandH


© 2016 The University of Sheffield
Agent Interacting with Agent
AGENT (a_1)
Intentions
Search
Search
g_{a1}
$\hat{h}_{a2}$
Actions
AGENT (a_2)
Search
Intentions
Search
h_{a2}
$\hat{g}_{a1}$
Agents exploiting knowledge of themselves to manipulate/ interpret each other
Moore, R. K. (2014). Spoken language Processing: Time to Look Outside? 2nd International Conference on Statistical Language and Speech Processing (SLSP 2014). Grenoble.
The University Of Sheffield.
UCREL, Lancaster
21st January2016

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*

The University Of Sheffield.

SPandH

Speech Energetics

Robin Hofe

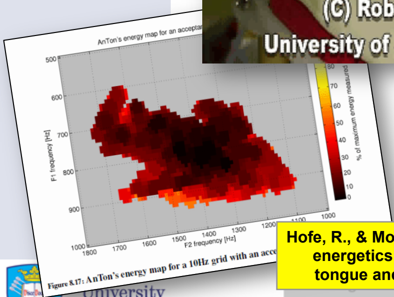
Hofe, R., & Moore, R. K. (2008). Towards an investigation of speech energetics using 'AnTon': an animatronic model of a human tongue and vocal tract. *Connection Science*, 20(4), 319–336.

The University Of Sheffield.

SPandH

‘AnTon’ – Animatronic Tongue

Robin Hofe



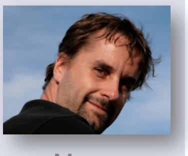
Hofe, R., & Moore, R. K. (2008). Towards an investigation of speech energetics using 'AnTon': an animatronic model of a human tongue and vocal tract. *Connection Science*, 20(4), 319–336.

The University Of Sheffield.

SPandH

© 2016 The University of Sheffield

Reactive Speech Synthesis

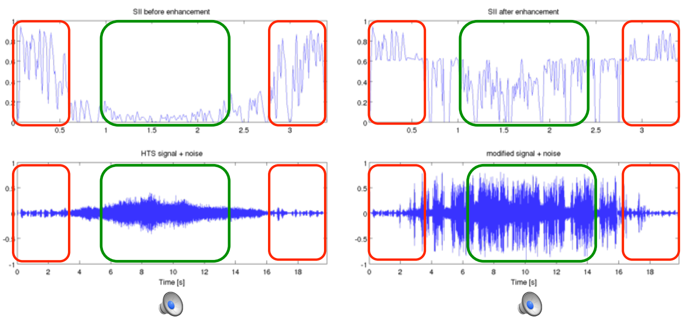


Mauro Nicolao

SCALE

Traditional TTS

Reactive TTS



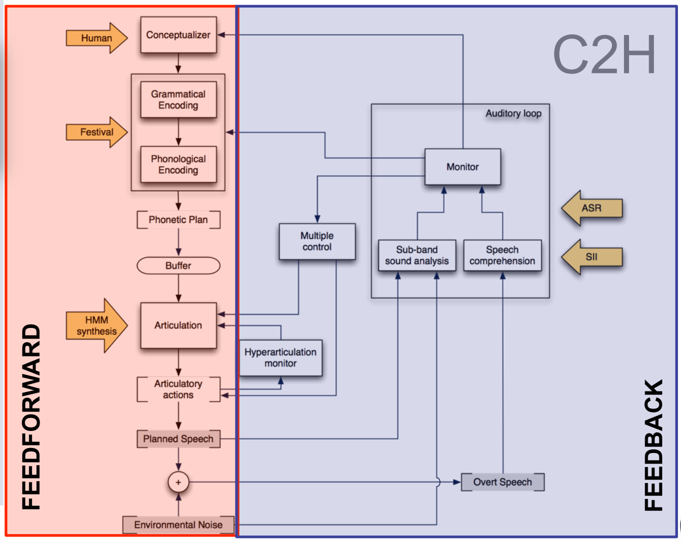
Moore, R. K., & Nicolao, M. (2011). Reactive speech synthesis: actively managing phonetic contrast along an H&H continuum, *17th International Congress of Phonetics Sciences (ICPhS)*. Hong Kong.

The University Of Sheffield.

SPandH

Computational H&H

Mauro Nicolao

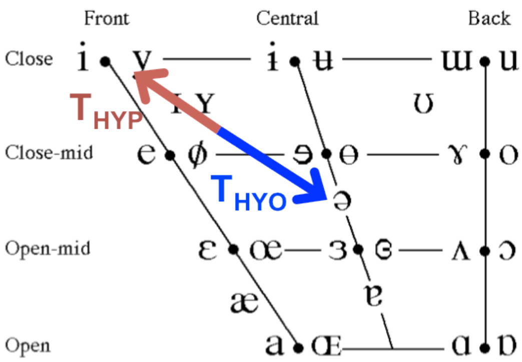


The University Of Sheffield.

SPandH

VPC: Vowel Production Control

Mauro Nicolao



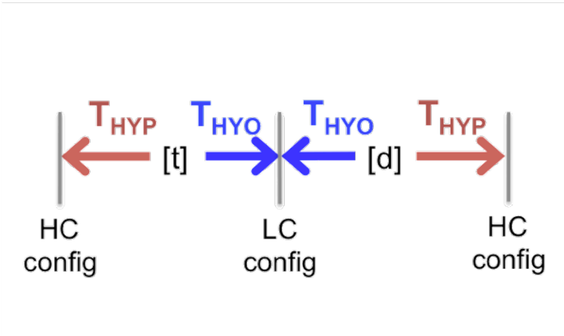
Nicolao, M., Latorre, J., & Moore, R. K. (2012). C2H: A computational model of H&H-based phonetic contrast in synthetic speech. *INTERSPEECH*. Portland, USA.

The University Of Sheffield.

SPandH

CPC: Consonant Production Control

Mauro Nicolao

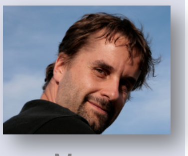


Nicolao, M., Latorre, J., & Moore, R. K. (2012). C2H: A computational model of H&H-based phonetic contrast in synthetic speech. *INTERSPEECH*. Portland, USA.

The University Of Sheffield.

SPandH

C2H: Experimental Setup



Mauro Nicolao

- 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

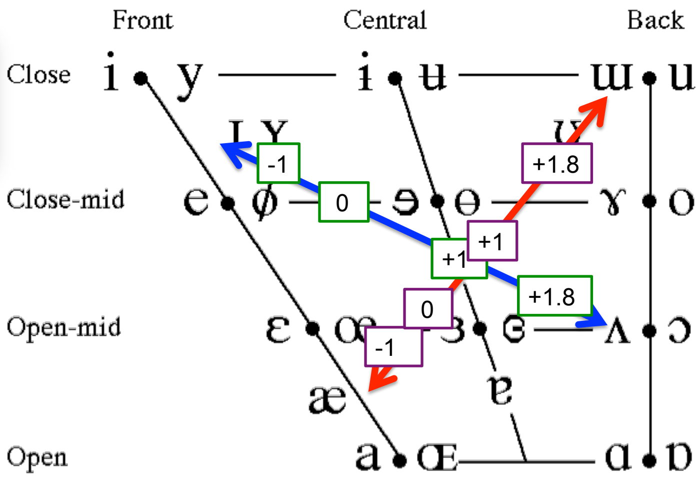
Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11), 1039–1064.

The University Of Sheffield.

SPandH

Mauro Nicolao

"I will say sat again"



The University Of Sheffield.

SPandH

VPC Results

	HYPO	NORM	HYPER
Mean Word Duration (s)	0.27	0.32	0.36
Mean Sentence Dur. (s)	2.98	3.50	3.91
Pause Duration (s)	0.13	0.15	0.17
LTAS 1-3 (dB SPL)	33.6	36.2	41.1
Spectral Tilt (dB/dec)	-6.2	-5.8	-4.7
Spectral CoG (Hz)	712	821	1024
F0 (Hz)	172.6	174.1	174.7
F0 range (Hz)	146-185	151-183	145-190
F1F2 area (Hz^2)	1014	29021	70509

Nicolao, M., & Moore, R. K. (2012). Actively managing phonetic contrast along an H&H continuum in automatic speech synthesis. *5th Workshop on Speech in Noise: Intelligibility and Quality*. Vitoria, Spain.

The University Of Sheffield.

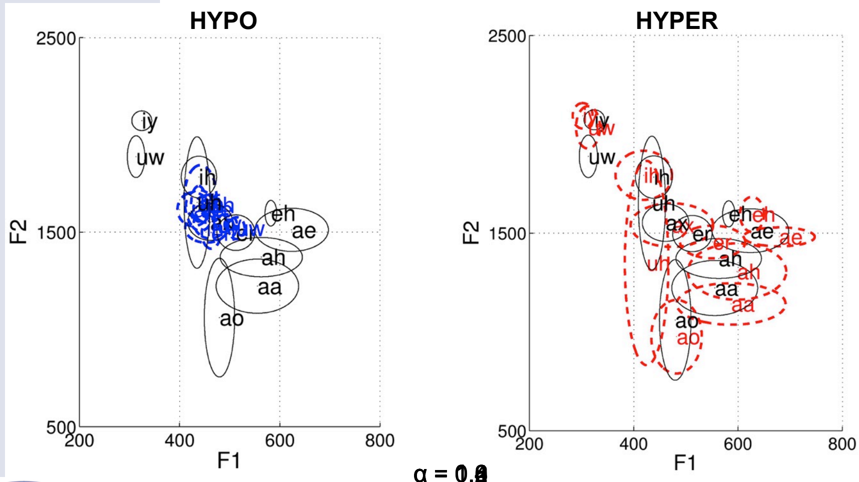
SPandH

CPC Results

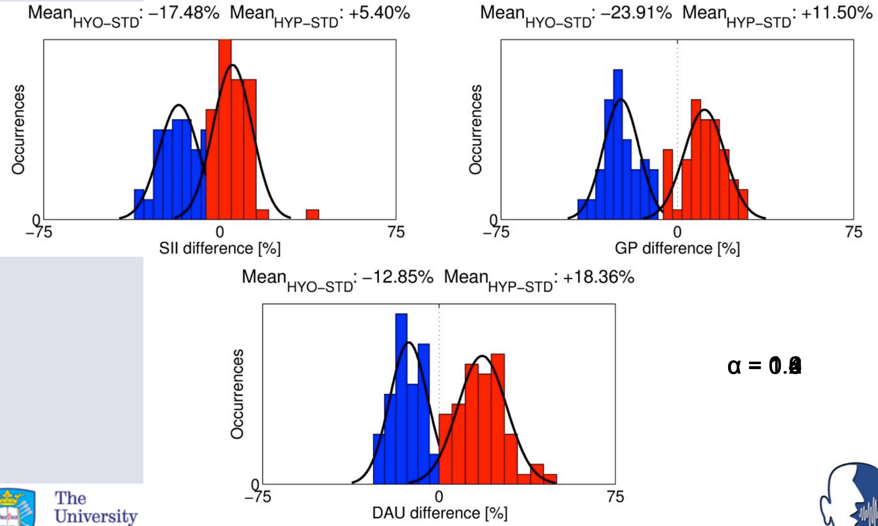
	HYPO	NORM	HYPER
Mean Word Duration (s)	0.31	0.32	0.33
Mean Sentence Dur. (s)	3.43	3.50	3.60
Pause Duration (s)	0.14	0.15	0.16
LTAS 1-3 (dB SPL)	35.4	36.2	38.4
Spectral Tilt (dB/dec)	-6.1	-5.8	-5.1
Spectral CoG (Hz)	547	821	1156
F0 (Hz)	174.1	174.1	173.4
F0 range (Hz)	144-185	151-183	150-184
F1F2 area (Hz^2)	41824	29021	56103

Nicolao, M., & Moore, R. K. (2012). Actively managing phonetic contrast along an H&H continuum in automatic speech synthesis. *5th Workshop on Speech in Noise: Intelligibility and Quality*. Vitoria, Spain.

Effect on Vowel Space



Effect on Intelligibility



α = 0.[illegible]

Example Speech: English Male

Type of noise	HYPO	NORM	HYPER
Speech Shaped Noise (SNR = 1 dB)	●	●	●
Competing Talker (SNR = -7 dB)	●	●	●
Clean	○	○	○

"The box was thrown beside the parked truck"

Example Speech: Italian Female

Type of noise	HYPO	NORM	HYPER
Car Noise (SNR = -4 dB)			
Babble Noise (SNR = -4 dB)			
Competing Talkers (SNR = -4 dB)			
Clean			

"Ti è mai successo di rimanere senza fiato?"

The University Of Sheffield.

39

SPandH

Example Speech: Italian Male

Type of noise	HYPO	NORM	HYPER
Car Noise (SNR = -4 dB)			
Babble Noise (SNR = -4 dB)			
Competing Talkers (SNR = -4 dB)			
Clean			

Nicolao, M., Tesser, F., & Moore, R. K. (2013). A phonetic-contrast motivated adaptation to control the degree-of-articulation on Italian HMM-based synthetic voices. In *8th ISCA Speech Synthesis Workshop (SSW8)*. Barcelona, Spain.

The University Of Sheffield.

40

SPandH

And Finally …

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

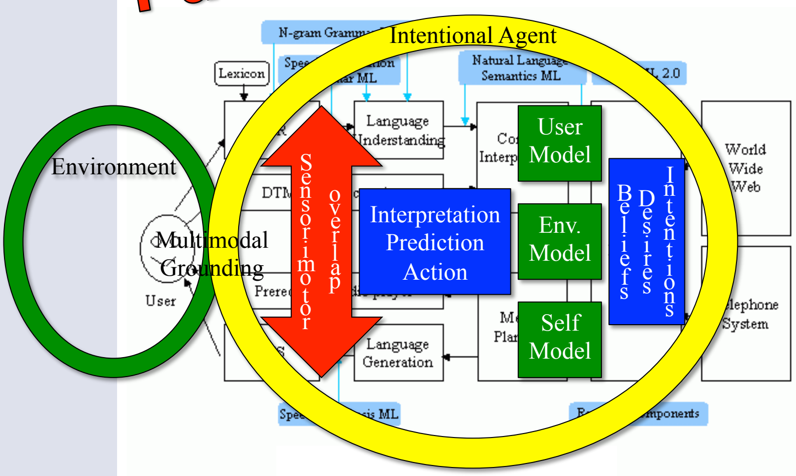
No? So why do we put people in a recording booth and ask them to speak?

In both cases the subject is obliged to <u>imagine</u> a crucial conditioning aspect of their behaviour

An appropriate <u>interactive</u> experimental methodology is the key to future progress

'Traditional' Architecture

Future



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.