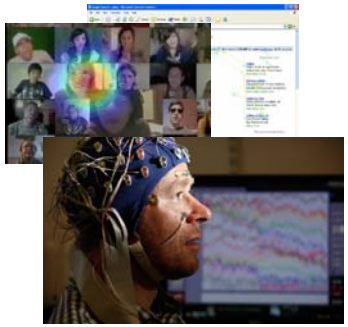
case studies in design informatics

Lecture 3:Speech and Dialogue Systems



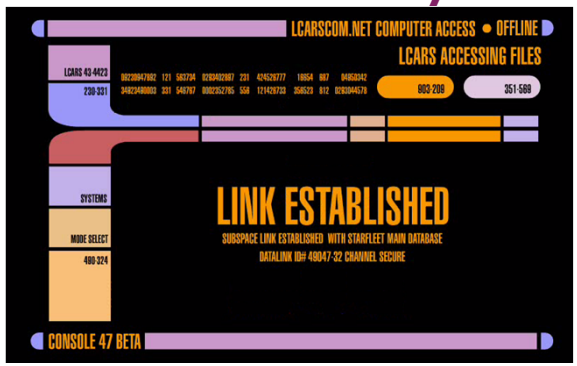
design informatics

NRlabs neuropolitics research

Robin Hill

Institute for Language, Cognition & Computation,
School of Informatics
and
Neuropolitics Research Lab,
Politics and I.R.
www.robin.org.uk
r.l.hill@ed.ac.uk

LCARS: Library Computer Access/Retrieval System



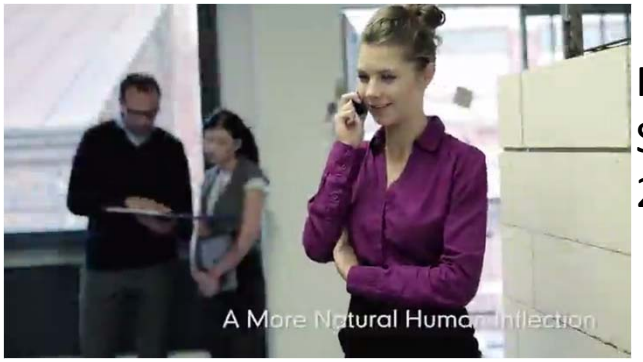
Sci-Fi but only just.

Ubiquitous computers and androids

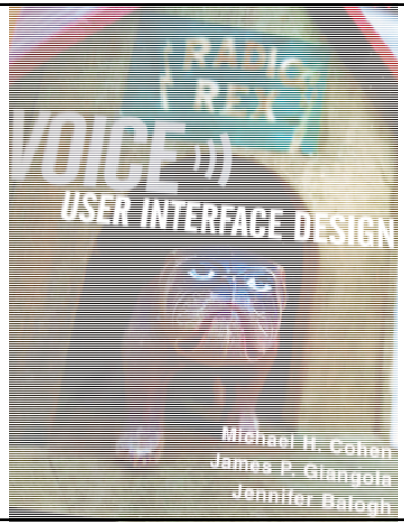
Sci-Fi but getting there.

WaveNet:

a generative model for raw audio



Real! Sept. 2016



Possibly the first ASR system was **Radio Rex**: a toy dog who sits in his house until the sound of his name calls him out. Produced in 1911.

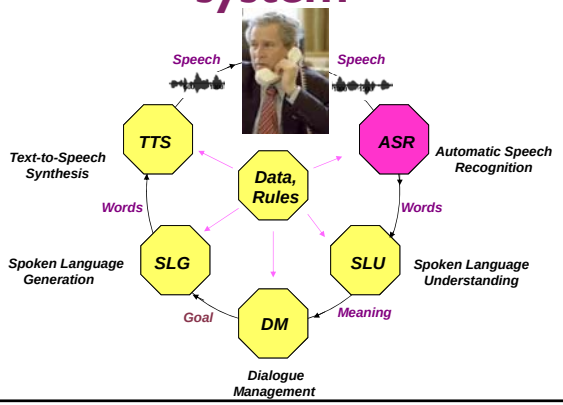
Spoken HCI & HRI

- Automatic Speech Recognition (ASR).
- Natural Language Processing (NLP):
  - formalising input (human to machine);
  - generating output (machine to human).
- Text-to-Speech (TTS) and speech synthesis [Dialogue].

Speech → text → speech

E.g. Neto, A. T., Fortes, R. P. M., & da Silva, A. G. (2008). Multimodal interfaces design issues: the fusion of well-designed voice and graphical user interfaces. In C. J. Costa, A. Protopsaltis, M. Aparicio & H. Oneill (Eds.), *Sigdoc'08: Proceedings of the 26th ACM International Conference on Design of Communication (pp. 277-278). New York: Assoc Computing Machinery.*

Anatomy of a spoken dialogue system



Basic architecture

1. Endpointing:
   - Detect onset and offset of user speech (slice up waveform).
2. Feature extraction:
   - Transform waveform into sequence of feature vectors (e.g. amount of energy at various frequencies), one vector per time period (e.g. 10ms).
3. Recognition:
   - Match feature vector sequence against most likely word sequence.
4. Natural language understanding
   - Extract meaning from word sequence (e.g. fill slots with values).
5. Dialogue management:
   - Select next system action (e.g. check database; speak to user; book flight).

From: Cohen, Michael H., Giangola, James P., & Balogh, Jennifer (2004). *Voice User Interface Design. Addison Wesley.*

Cohen, Giangola & Balogh (2004)

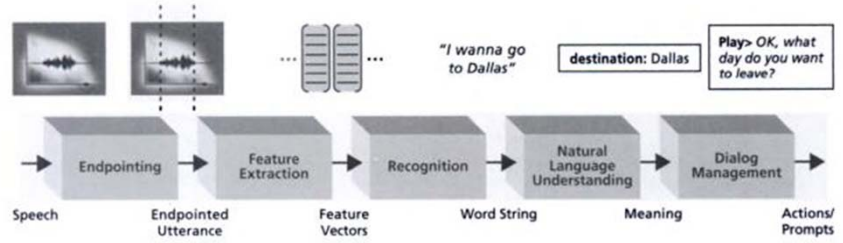


FIGURE 2-7 *The processing sequence for handling one spoken input from a caller.*

Traditional dialogue management

- State-based systems
  - Rigid and unadaptive.
- Frame-based systems
  - Fixed domain of discourse; highly scripted.
- Plan-based, or agent-based, systems
  - More flexible and generalisable.
  - Attempts to cope with ambiguity and error handling (more than "if not then fail" approach).

State-based systems

- Generate utterances and recognize users' responses according to a state-transition model, like a flowchart.
- Simple and intuitive, easy to implement and often used in working systems.
- Fine with small, closed state spaces (i.e. limited options and paths).

Frame-based systems

- Fit users' responses into pre-defined slots in **frames** to estimate user goals.
- Often used for telephone-based dialogue systems, e.g. for providing weather or transportation information.
- Frame-based systems can handle more complex information, but involve more effort for preparation of such frames of knowledge.

Plan-based [agent-based] systems

- Use a set of rules to change the internal states of an agent to navigate through conversation.
- These can handle the most complex interactions, but require very advanced natural-language processing and well defined sets of rules.
- This approach is often used in research but rarely used in working systems.

DS issues / challenges

- Output will suffer if there is a fault in any part of the pipeline.
- Dialogue systems and voice interfaces **have** to work in real-time (i.e. human-time).
- Conversational dialogue (multiple turns) requires complex **discourse coherence**.
  - More than commands such as “Computer: open Word!”

Discourse coherence

- Discourse: set of multi-sentence linguistics units.
- Discourse coherence: the structure and meaning of discourses (covering monologues as well as dialogues).
  - Many important entailments and cross-references established.
  - **Lexical chains** are sets of the same or related words appearing in consecutive sentences.
  - Breaks in lexical chains often indicate topic shifts.

Automatic Speech Recognition requires

1. Acoustic models
   - Represent (statistically) how each phoneme (basic sound) may be pronounced.
2. Dictionary
   - Lists words with their (multiple) pronunciations: which acoustic models sequence into which word models.
3. Grammar
   - Defines all possible user inputs (word strings and meanings).
   - Rule based or statistical (SLMs are more flexible; require training).
4. Recognition model
   - A search to find best-matching path(s) through the lattice.
   - Returns a confidence measure (match score between input vectors and best path).

(Cohen et al., 2004)

Cohen, Giangola & Balogh (2004)

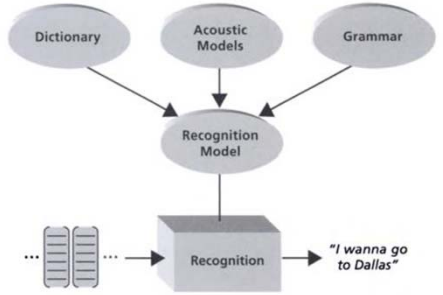
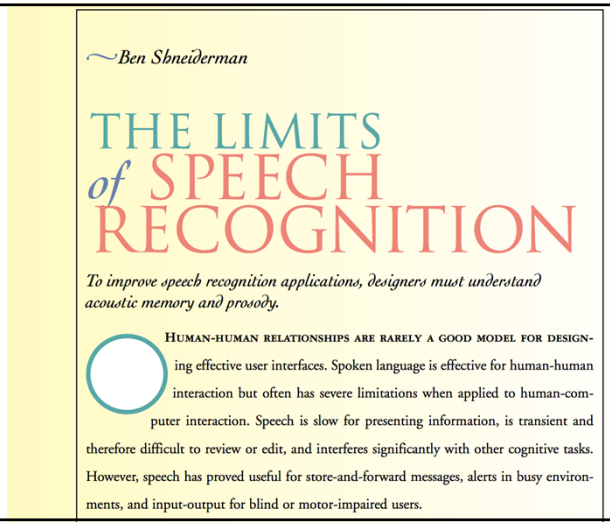


FIGURE 2-8 *The recognizer searches the recognition model to find the best-matching word string. The recognition model is built from the acoustic models, dictionary, and grammar.*



Ben Shneiderman

THE LIMITS *of* SPEECH RECOGNITION

To improve speech recognition applications, designers must understand acoustic memory and prosody.

HUMAN-HUMAN RELATIONSHIPS ARE RARELY A GOOD MODEL FOR DESIGNing effective user interfaces. Spoken language is effective for human-human interaction but often has severe limitations when applied to human-computer interaction. Speech is slow for presenting information, is transient and therefore difficult to review or edit, and interferes significantly with other cognitive tasks. However, speech has proved useful for store-and-forward messages, alerts in busy environments, and input-output for blind or motor-impaired users.

Ben Shneiderman

- Renowned HCI expert.
- Gave a lecture in Informatics in June coinciding with his new book "The New ABCs of Research: Achieving Breakthrough Collaborations" (Oxford University Press).
- Sceptical of generic speech interfaces. Will discuss the following paper here:

  Shneiderman, Ben (2000). The limits of speech recognition. *Communications of the ACM, 43(9), 63-65. doi: 10.1145/348941.348990.*

Concept

- **Human-human relationships are rarely a good model for designing effective user interfaces.**
- Spoken language is effective for human-human interaction but often has severe limitations when applied to human-computer interaction.
- Speech is slow for presenting information, is transient and therefore difficult to review or edit, and interferes significantly with other cognitive tasks.
- However, speech has proved useful for store-and-forward messages, alerts in busy environments, and input-output for blind or motor-impaired users.

(Shneiderman, 2000)

Niche applications

- Speech recognition and generation is sometimes helpful for environments that are hands-busy, eyes-busy, mobility-required, or hostile and shows promise for telephone-based services.
  - Telephone-based speech-recognition applications, such as voice dialling, directory search, banking, and airline reservations.
    - May be useful complements to graphical user interfaces.
  - Dictation input is increasingly accurate, but adoption outside the disabled-user community has been slow compared to visual interfaces.
    - Obvious physical problems include fatigue from speaking continuously and the disruption in an office filled with people speaking.

(Shneiderman, 2000)

Cognitive processes

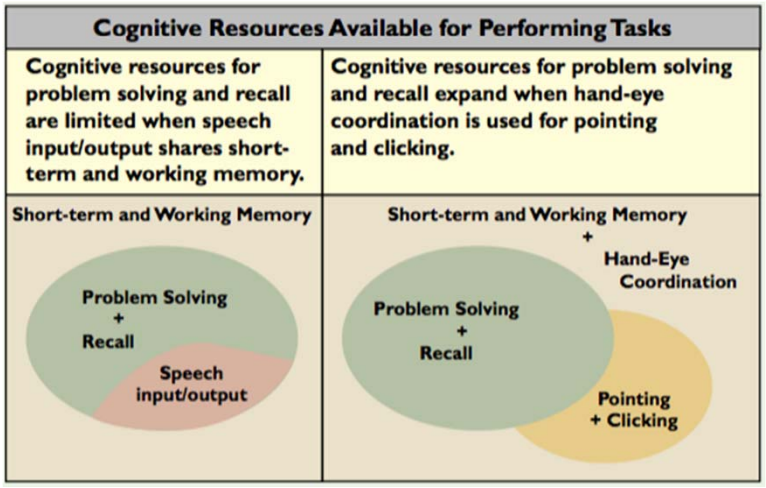
- By understanding the cognitive processes surrounding human "acoustic memory" and processing:
  - interface designers may be able to integrate speech more effectively and guide users more successfully.
- By appreciating the differences between human-human interaction and human-computer interaction:
  - interface designers may then be able to choose appropriate applications for human use of speech with computers.

(Shneiderman, 2000)

Human acoustic memory

- Categorised as short-term and working memory.
- The part of the human brain that transiently holds chunks of information and solves problems also supports speaking and listening.
- Therefore, working on tough problems is best done in quiet environments:
  - without speaking or listening to someone.
- However, because physical activity is handled in another part of the brain, problem solving is compatible with routine physical activities like walking and driving.

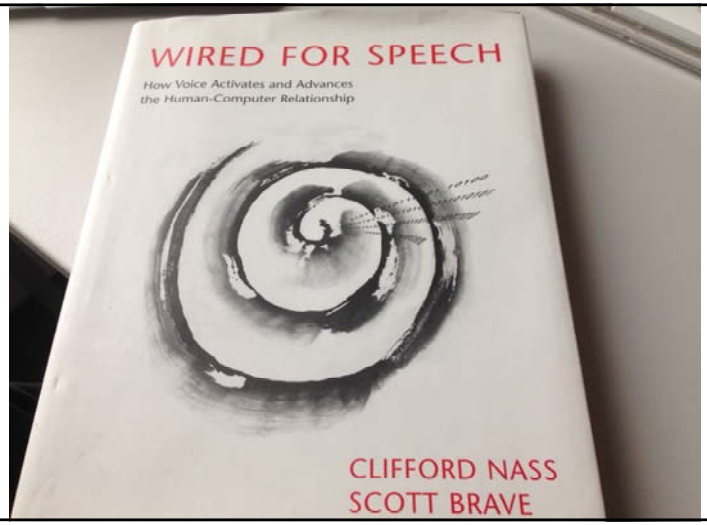
(Shneiderman, 2000)



Neural factors

- Humans speak and walk easily but find it more difficult to speak and think at the same time.
  - Similarly when operating a computer, most humans type (or move a mouse) and think but find it more difficult to speak and think at the same time.
- Hand-eye coordination (motor control) uses a different brain region, so typing or mouse movement can be performed in parallel with problem solving.
  - Speaking consumes precious cognitive resources, inhibiting simultaneous problem solving.
  - Proficient keyboard users can have higher levels of parallelism in problem solving while performing data entry.
- This may explain why after 30 years of ambitious attempts to provide military pilots with speech recognition in cockpits, aircraft designers persist in using hand-input devices and visual displays.

(Shneiderman, 2000)



More optimistic view?

- Massive growth in mobile technology makes speech an obvious choice.
- Nass, Clifford, & Brave, Scott (2005). *Wired for Speech: How Voice Activates and Advances the Human–Computer Relationship.* Cambridge, MA: MIT Press.

Nass & Brave, 2005

- Humans are experts at extracting the social aspects of speech.
- Gender, personality, emotion:
  - who to like, who to trust, who to do business with.
- How do we deal with technologies that talk or listen?
  - The brain rarely distinguishes between speaking to a machine and speaking to a person.
  - In fact, the psychology of interface speech is just the psychology of human speech.
  - "Voice interfaces are intrinsically social interfaces."
- Designers can use our powerful responses, to
  - "increase liking, trust, efficiency, learning, and even buying" (p4)

Gender of voices

- Technologies do not really have gender.
- But from an early age, in all cultures, people spend most of their time with people of the same gender.
  - 6-month old babies can categorize voices into female or male.
  - Adults categorize a voice within seconds.
- Some differences are cultural:
  - e.g. In USA, compared to men, women's voices have wider pitch range, more expressiveness, more utterances with rising pitch.
  - Their speech also includes more questions, and more social-relational information.

(Nass & Brave, 2005)

Gender and similarity

- Gender categorization enables social identification.
  - Is this person is the same group as me, or not?
- Because similarity encourages positive feelings, social identification leads to:
  - Greater trust, liking, perceived intelligence, etc.

(Nass & Brave, 2005)

Similarity and social alignment

- Similarity increases:
  - predictability and hence perceived safety
  - understandability and hence cognitive economy
  - (in the past) social support
  - Familiarity; liking
- Synthetic speech maybe quite unlike human speech (random pauses, misplaced accents, glitches, etc.) but people are still good at perceiving male, female or non-human voices.
  - Perhaps this triggers social identification.
- Danger of reinforcing stereotypes (even in a digital world).

(Nass & Brave, 2005)

Human-computer success and failure

- In cases of success:
  - Computer can complement User (kindness).
  - Computer can thank User (politeness).
  - Computer can spread the praise (credit sharing).
  - Computer can acknowledge praise from User (politeness).
- But failure is frequent for VUIs:
  - 95% accuracy means 1 in 20 words misunderstood.
  - When that happens, the system may not know what to do next.
  - So, it must acknowledge the problem.

(Nass & Brave, 2005)

Strategies to cope with failure

If it follows human precedent, there are two choices for the dialogue system:

1. Take the blame:
   - Sorry, I didn't understand you.
   - Sorry, I didn't get that.
   - I'm sorry, I'm not finding a match.
2. Blame the partner:
   - The system could not recognise what you said.
   - Please speak more clearly.
   - Please pay attention.

(Nass & Brave, 2005)

Error avoidance / minimisation

- **Lexical alignment**:
  - if system says: "read, save, or throw away", users are much less likely to say "delete". Adopt and share a common language [lexical entrainment and conceptual pacts].
- If system can recognise full phrases (instead of single words), users get a greater sense of control. But recognition rates tend to go down.
  - Nonetheless **syntactic alignment** means that if the system prefers to use syntactic structures that are easy for it to recognise, users will tend to use them too.
- Similarly, if recognition is better at a particular speech rate, the system can entrain the user to speak at that rate by using it itself.

(Nass & Brave, 2005)

Expectation and bonding

- Speaking less competently:
  - will move users towards simpler (and easier to recognise) speech.
  - E.g. synthetic vs natural voice; slow speech rate; non-grammatical utterances, short words, simple sentences.
- Users will try harder to work with systems that they identify with.
  - So, match gender, personality, accent.
- Reciprocity helps:
  - if the system repeats words specific to the user's vocabulary, it suggests that it is paying full attention (and thus committed to the success of the interaction).
  - Pausing also suggests that care has gone into response construction.

(Nass & Brave, 2005)

Tesco self-service

- Last year Tesco replaced the audio voice that talks to customers at its self-service and scan-as-you-shop checkouts.
- The grocery giant has described the new voice as "friendlier, more helpful and less talkative" and said the change came about following complaints from customers that the previous instructor's voice was "shouty" and "irritating".
- Twitter users were quick to notice the most obvious difference between the two voices: the annoying voice is female and the friendlier new voice is male.

http://www.telegraph.co.uk/finance/newsbysector/epic/tsco/11772689/Tesco-replaces-irritating-unexpected-item-in-the-bagging-area-self-checkout-voice-with-male-recording.html

JAST

Arbitrary initial object layout

Resulting objects after completion of multimodal dialogue sequence

Pre-order description strategy, basic reference strategy

SYSTEM First we will build a windmill. Okay?

USER Okay.

SYSTEM To make a windmill, we must make a snowman.

SYSTEM *[picking up and holding out red cube]* To make a snowman, insert the green bolt through the end of the red cube and screw it into the blue cube.

USER *[takes cube, performs action]* Okay.

Post-order description strategy, full reference strategy

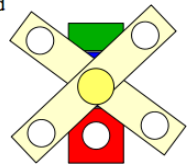
SYSTEM First we will build a windmill. Okay?

USER Okay.

SYSTEM *[picking up and holding out red cube]* Insert the green bolt through the end of this red cube and screw it into the blue cube.

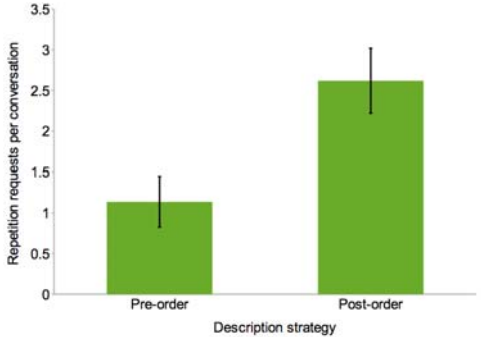
USER *[takes cube, performs action]* Okay.

SYSTEM Well done. You have made a snowman.



(Foster et al., 2009)

Robots should explain things first

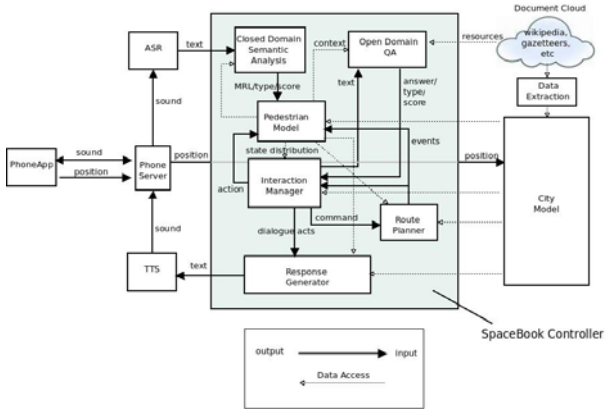


Foster, M.E., Giuliani, M., Isard, A., Matheson, C., Oberlander, J. and Knoll, A. [2009] Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09), pp1818--1823. Pasadena, California, July 2009.*

SpaceBook ideal example

- User: So, where am I?
- EARS: You are on the Royal Mile, not far from your hotel.
- User: OK… so what's happening in Edinburgh today?
- EARS: Well, most of the tourist attractions are open as usual.. The galleries, museums, cinema, shops… what time do you have available?
- User: Oh, all day.
- EARS: The weather's nice – how about a walk in the Botanics.
- User: Nah – I hate walking. Right now I'm feeling thirsty.
- EARS: There is a nice café just around the corner from where you are.
- User: Sounds good
- EARS: Turn to your left, walk straight ahead 100m towards the large glass fronted building you can see in front of you. Can you see it?
- User: Yep.
- < … >
- EARS: OK, turn right, you probably recognise this street - the café is on the opposite side of the road – called 'The Late Latte' – can you see it?
- User: Yep. So what happens in that large glass fronted building?
- EARS: It's the festival theatre. Did you want to know what shows are on just now?
- User: OK – yeh. I might take a look inside.
- EARS: Theres a café inside the festival theatre too.
- User: Sounds good to me.

Overall architecture



Example communicative functions

CF	Example	
instruct	Turn left here.	instruct(turn(rdir=left))
request	Could you repeat that please?	request(repeat(X), prevUtt(X))
suggest	How about visiting the Royal Museum?	suggest(visit(X), museum(X))
inform	You are now on Princes Street.	inform(User(9999),InOn(9999,X), Street(X), IsNamed(X, 'Princes Street'))
propositionalQuestion	Can you see the station?	propositionalQuestion(inViewShed(X), museum(X))
	Do you mean the Royal Museum?	propositionalQuestion(mean(X), isNamed(X,"Royal Museum"))
setQuestion	Where are you now?	setQuestion(?X.location(X), currentLocation(X))
	Which museum do you mean?	setQuestion(?X.museum(X), mean(addressee,X))
autoPositive	Okay.	autoPositive()
autoNegative	Sorry?	autoNegative()
pausing	Just a moment.	pausing()
initialGreeting	Hello!	initialGreeting()
returnGreeting	Hi!	returnGreeting()
apology	I'm sorry.	apology()
acceptApology	That's alright.	acceptApology()
thanking	Thank you.	thanking()
acceptThanking	You're welcome.	acceptThanking()
initialGoodbye	Goodbye.	initialGoodbye()
returnGoodbye	Bye.	returnGoodbye()

SpaceBook vocabulary

- Attempted to limit vocabulary (minimise possible state space) to speed up ASR identification.
  - Increased problems.
  - People could ask questions about anything.
  - People would chat about anything they happened to see.
- Navigation uses a lot of place names. A problem for both ASR and TTS. Try pronouncing the following local roads (or transcribing the speech):
  - Buccleuch Place.
  - Cockburn Street.

CereProc (Informatics success story)

- CereProc offers a range of voices in many accents. We can create amazing new voices quickly due to our innovative voice creation system. Many of our voices are built exclusively for specific customers and applications.
- Based in Edinburgh, we are very proud of our Scottish voices. The Scottish accent has achieved the highest level of acceptability across a range of accredited surveys. This is one reason Scotland has long been the location of choice for contact centre operations.

https://www.cereproc.com/en/products/voices

hitchBOT (Ryerson University)

A child-sized robot is currently hitch-hiking its way across Canada - relying on the kindness of strangers and its own powers of charm. Well, its built-in speech recognition and processing capabilities anyway.

https://www.indy100.com/article/this-robot-is-trying-to-hitchhike-across-north-america-by-itself--gyE3ktTffe