

# Human-Robot Interaction

## Module 4: Interaction

### Lecture 4: Dialogue management; speech production

David Vernon  
Carnegie Mellon University Africa

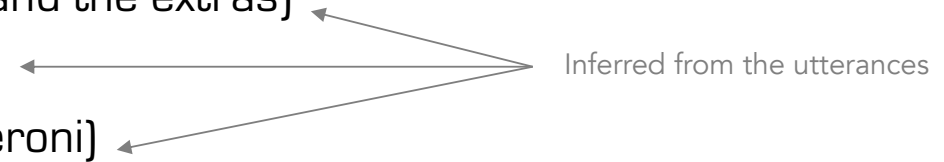
[www.vernon.eu](http://www.vernon.eu)

# Topics

- Dialogue management
  - Basic principles
  - Practice in HRI
    - Turn-taking in HRI
    - The role of timing
- Speech production
  - TTS engines
  - Chatbots & cloud services

# Dialogue Management

- Dialogue management (DM) is the process that keeps a conversation flowing
  - Written
  - Spoken
- Speech Keep track of the state of the conversation
  - What the robot needs to know (e.g. the order and the extras)
  - What the robot knows (e.g. the order: pizza)
  - What the robot has yet to establish (e.g. pepperoni)



# Basic Principles

## Range of complexity of DMs


- Strict ordering of dialogue
  - Closed and well-contextualized tasks
    - Registering guests
    - Filling out forms
    - Taking orders
  - System initiative DMs
    - No latitude for the human to change the course of the dialogue
- User-initiative DMs
  - The user takes the lead
  - The DM intervenes when more information is needed
- Mixed initiative DM
  - Combination of system initiative and user-initiative approaches



This is the way it's presented in the book:  
The sub-classification is not very clear

# Basic Principles

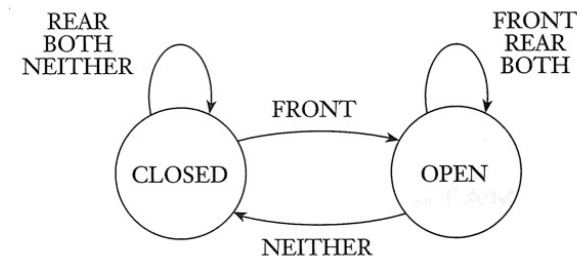
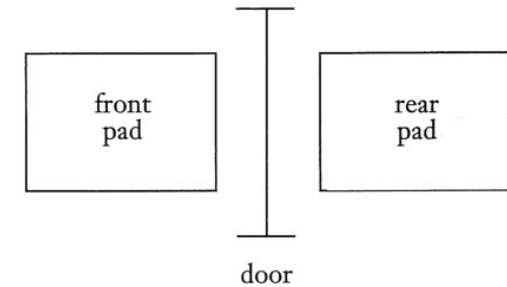
## Range of complexity of DMs

- System initiative DMs
    - Strict ordering of dialogue
    - Closed and well-contextualized tasks
      - Registering guests
      - Filling out forms
      - Taking orders
      - No latitude for the human to change the course of the dialogue
  - User-initiative DMs
    - The user takes the lead
    - The DM intervenes when more information is needed
  - Mixed initiative DM
    - Combination of system initiative and user-initiative approaches
- 
- This might be better

# Basic Principles

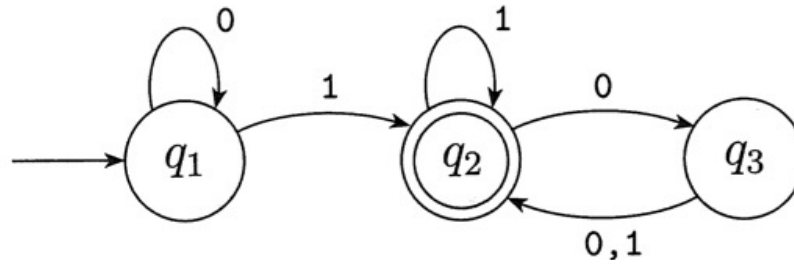
Simple DMs are finite state machines  
(FSMs)

- A set of rules that define
  - the different states of the system  
(i.e. the dialogue)
  - The conditions governing the change  
from one state to another
- Strict control of flow



# Basic Principles

Also called **Finite Automata**



1.  $Q = \{q_1, q_2, q_3\}$ ,
2.  $\Sigma = \{0,1\}$ ,
3.  $\delta$  is described as

|       | 0     | 1     |
|-------|-------|-------|
| $q_1$ | $q_1$ | $q_2$ |
| $q_2$ | $q_3$ | $q_2$ |
| $q_3$ | $q_2$ | $q_2$ |

4.  $q_1$  is the start state, and
5.  $F = \{q_2\}$ .

$$\delta: Q \times \Sigma \rightarrow Q$$

# Basic Principles

More advanced DMs allow events to control the flow of the dialogue

- The control of flow can be interrupted
- "Non-linear" dialogue flows

e.g., human asks the time in the middle of ordering the pizza



# Basic Principles

## Use a planner

- Instead of writing explicit rules for all possible circumstances
- The planner determines the action (the questions) needed to complete any missing information
- Planners are also used to determine the actions required for a robot to achieve a goal

# Practice in HRI

- Companies that offer speech recognition services often offer
  - Dialogue management services
  - Speech production services
- The most popular DMs are event based
- But they are not suitable for free-flowing open discourse

# Turn-taking in HRI

Back-channeling: the responses given by a listener to indicate that she or he is still engaged

- Verbal cues ("really?")
- Non-verbal cues (nod your head)
- Some robots use visual backchanneling, e.g. changing the color of the area around the eyes
- Getting the timing right is difficult as it's dependent on the speaker

# The Role of Timing

Timing is critical in natural interaction

- Too slow and it's disturbing
- Too quick and it's insincere
- Yes/no answers require a faster response (100 ms)
- Answers requiring more thought take longer
- Sometimes, answers are given **before** the end of the question (prospection in dialogue)
- Typically, robots are slow compared to humans
- Just-in-time synthesis: start the articulation before the sentence is completely formulated
- Incremental spoken-dialogue: taking action before a spoken instruction is finished

# Speech Production

Converting a written response by the system to speed

- Speech synthesis / Text-to-speech (TTS)
- Primary approaches
  1. Concatenation
  2. Parametric
  3. Generative deep neural networks

# Speech Production

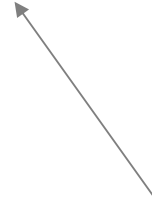
- **Concatenation**
  - Record an actor speaking
  - Assemble the phonemes to create the message (plus some "smoothing" between the phoneme)
  - Creates very natural speech
  - Not very flexible: new voices require new recordings
- **Parametric**
  - More flexible: allows customization of voice and prosody (rhythm, stress, intonation)
  - Classical approaches are not as natural
  - However, **generative deep neural networks**
    - Produce speech that is virtually indistinguishable from human speech
    - Used by Google as the voice of its digital assistant

# Speech Production Practice in HRI

- Simple TTS can run on robot hardware
- The most natural TTS systems use deep neural networks and are cloud-based
- TTS can also provide timing for phonemes which can aid synchronization with physical gestures
- The voice should fit the appearance of the robot
- The type of voice affects the social perception of social

# Speech Production Practice in HRI

"Robots with a male voice are anthropomorphized and evaluated more favorably by men than by women, and vice versa"



It is not clear what this means.

A robot with a female voice is evaluated more favourably by women?

Treat with caution: There is a great deal of research on the use of female vs male voices and the impact on gender bias.



# Speech Production Practice in HRI

- Adaptive prosody and emotion are not commonly available on TTS engines
- Synthesized voices don't adapt to the auditory context (quiet room vs loud exhibition hall)

# Speech Production Practice in HRI

Advanced chatbot technology

- Siri from Apple
- Cortana from Microsoft
- Alexa from Amazon
- Bixby from Samsung
- Google Assistant

is now being made available for developers

- Cloud Speech from Google
- Alex-based Cognitive Services from Amazon

# Speech Production Practice in HRI

This means you don't necessarily need to build your own software for

- Speech recognition
- Understanding, and
- Synthesis

# Speech Production Practice in HRI

Question: is the use of cloud services a valid model for HRI in Africa? Consider

- Cost
- Connectivity
- Lack of bias in training data sets

More generally, is the use of cloud services a valid model for AI & ML in Africa?

- IBM Cloud
- Amazon AWS
- Microsoft Azure
- Google Cloud

# Reading

Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., Sabanovic, S. (2020). Human-Robot Interaction - An Introduction, Cambridge University Press.

Chapter 6 – Verbal Interaction, pp. 106-113.