UiO : University of Oslo

INF5820 Part 2: Spoken Dialogue Systems

Pierre Lison,
Language Technology Group (LTG)
Department of Informatics

Fall 2014

Outline

- Practical details
- What is a spoken dialogue system?
- Architectural schema
- Hot research topics
- Summary

Outline

- **Practical details**
- What is a spoken dialogue system?
- Architectural schema
- Hot research topics
- Summary

Course objectives

1. Introduce the field of *spoken dialogue systems* (SDS) and its applications
2. Understand the main features of spoken dialogue, and why it can be difficult to process
3. Describe the core components of dialogue system architectures
4. Explain how dialogue systems are practically designed, built and evaluated

Schedule

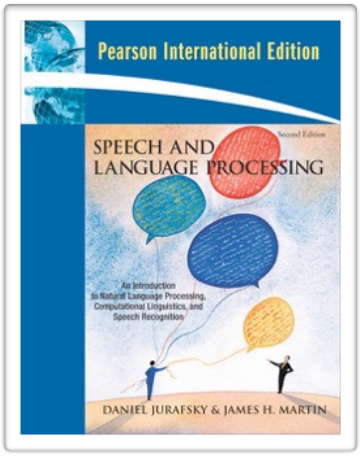
Fre.	24.10	Introduction, architectures, current research
Fre.	31.10	What is spoken dialogue?
Fre	07.11	Phonetics and speech recognition
Fre	14.11	Natural language understanding
Fre	21.11	Dialogue management
Fre	28.11	Generation, speech synthesis, evaluation
Fre	*5.12*	*Summary and Q&A*
Ma	*8.12*	*Written exam (start at 14:30)*
Ons.	29.10	Presentation of the project
Ons	5.11	Discussion of paper, exercises on spoken dialogue
Ons **9:00!!!**	12.11	Exercises in phonetics, help with project
Ons	19.11	Last minute help with assignment **Project deadline!**
Ons.	26.11	Presentation of projects

Assignment

- One obligatory assignment: a project where you will develop your own spoken dialogue system!
  - Submission deadline: **November 19** (23:59)
  - Presentation of the projects on November 26
  - More information on this project will be given during the first gruppetime
  - Submission through Devilry (grades: pass/fail)

Pensum

- Main reference material: **slides** from the lectures
- Jurafsky & Martin's «*Speech and Language Processing*», 2nd ed.
  - The selection of relevant chapters and sections will be provided in due time
- Studying the relevant book chapters and the slides should make you ready for the exam



A few words about myself

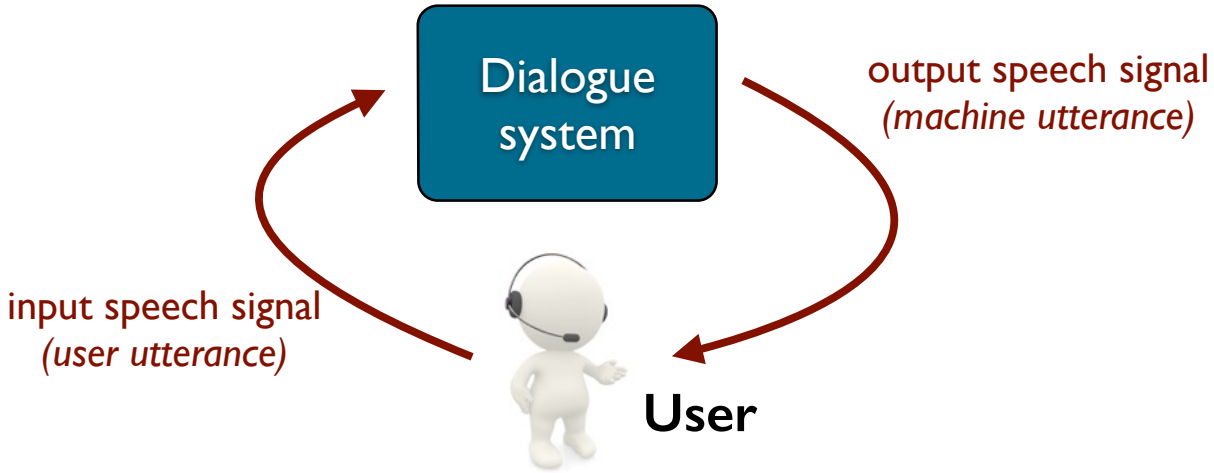
- Postdoctoral Research Fellow in the *Language Technology Group (LTG)*
- PhD from UiO (2014) with a thesis on dialogue management
- Before coming to UiO: researcher at the DFKI (Germany), working on human-robot interaction
- Education:
  - Computer Science & Engineering from University of Louvain (BE, 2006)
  - M.Sc. in Language Science & Technology from University of Saarland (DE, 2008)

Outline

- Practical details
- **What is a spoken dialogue system?**
- Architectural schema
- Hot research topics
- Summary

Spoken dialogue systems?

A spoken dialogue system (SDS) is a computer system designed to interact with humans using *(spoken) natural language*



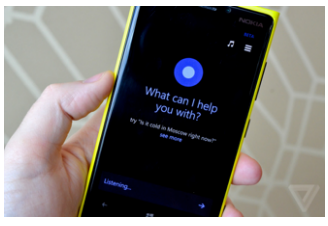
What for?

- Why could it be useful to interact with a machine using natural, spoken language?
  - Very intuitive interface, with virtually no need for training or expertise: all you need is to talk!
  - Touch-based interfaces can sometimes be inadequate (mobile phones) or dangerous (car driving)
  - Language is the ideal medium to express complex ideas in a flexible and efficient way

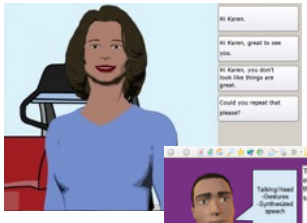
Examples of applications

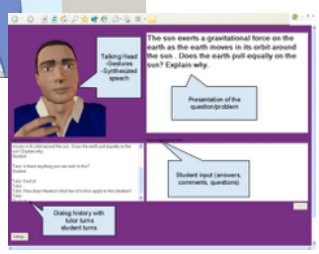
Voice-based access to information and services (orders, transport, support, etc.)

Mobile virtual assistants (Siri, Cortana, Google Now, etc.)



In-car navigation & control systems





Tutoring systems and healthcare assistants

Smart home environments

... and service robots

Motivation

- Why study spoken dialogue systems?
- A few practical reasons:
  1. Ultimate «intuitive» human-machine interface: you only need to be able to speak!
  2. R&D investment from major companies (Google, Apple, Microsoft, Nuance, Amazon, AT&T, Honda)
  3. Rapidly developing research area, huge potential in many domains (mobile apps, assistive technologies, robots)

Motivation

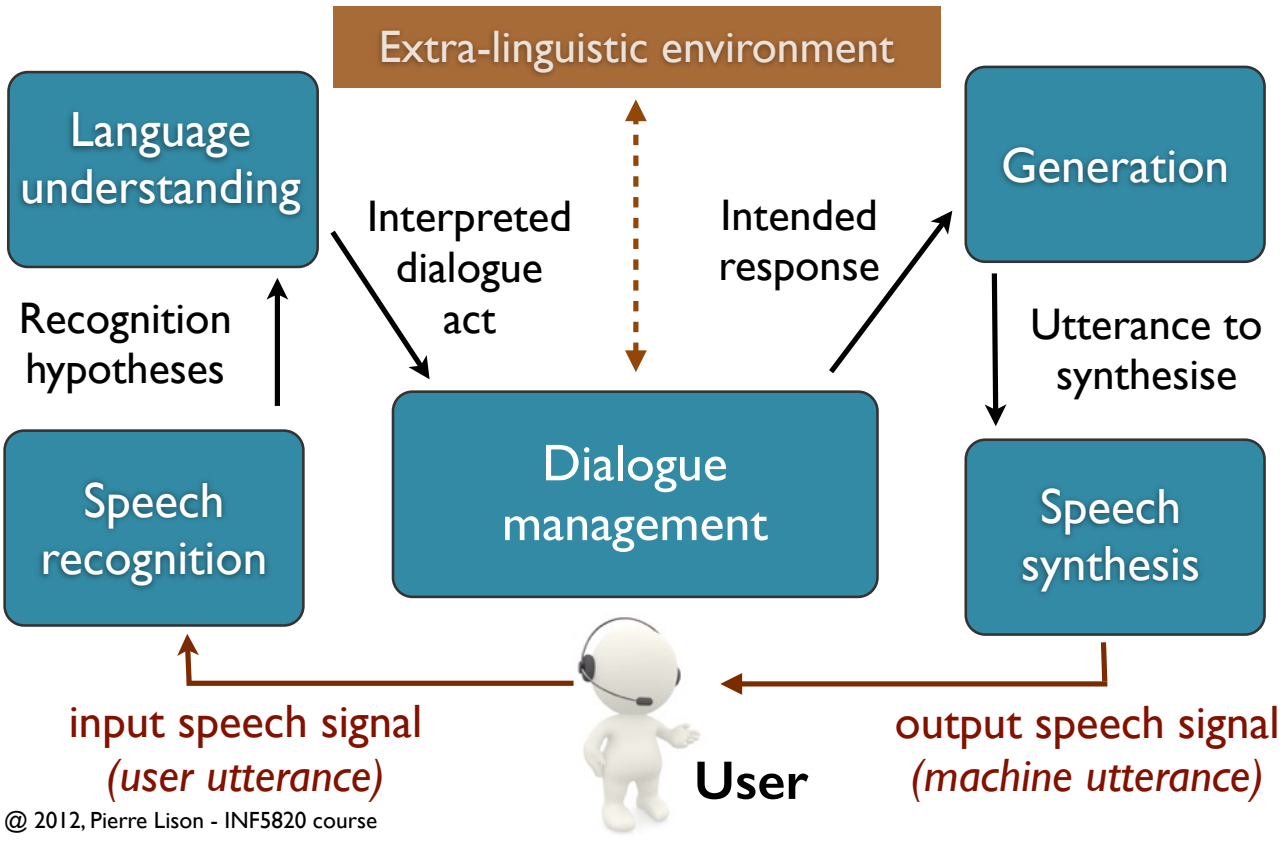
- And some theoretical reasons:
  1. Study language «*in the wild*», in the context of real interactions
  2. Study language *as a whole*, all the way from speech signals to high-level representations, and back
  3. Playground for many A.I. techniques: need to *sense*, *reason*, and *plan* under *uncertainty*, in *real-time*, with *several agents*... and continuously *learn* and *adapt* from *experience*

vs.

Outline

- Practical details
- What is a spoken dialogue system?
- **Architectural schema**
- Hot research topics
- Summary

A simple schema



Automatic speech recognition

- *Automatic speech recognition* (ASR) converts the speech signal into a list of hypotheses about what the user said
- This list is often called the N-best list, and is typically accompanied by some kind of confidence scores

47.6	this legend on spoken data system it really interesting dont you sink
45.2	this lecture on spokane log systems is really interesting doesn't using
38.9	these lectures on *UNK* dialogue systems is really interesting doesn't think

Language understanding

- *Natural language understanding* (NLU) covers a range of processing tasks responsible for extracting the *meaning* of a given user utterance
- Can include a semantic parser, a reference resolution engine, disfluency and error correction tools, etc.

47.6	this legend on spoken data system it really interesting dont you sink
45.2	this lecture on spokane log systems is really interesting doesn't using
38.9	these lectures on *UNK* dialogue systems is really interesting doesn't think

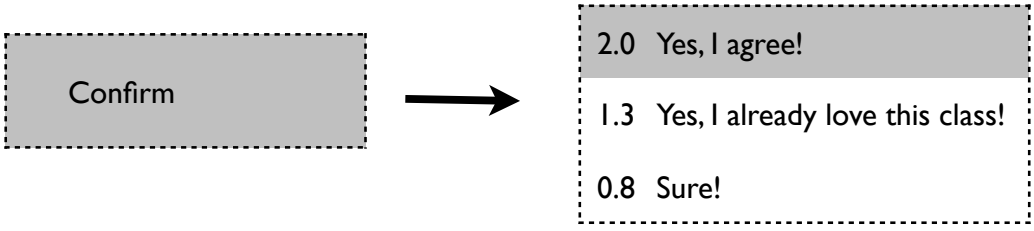
0.7	$Assert(lecture, interesting)$ $\wedge\ Topic(lecture, SDS)$ $\wedge\ FeedbackRequest$
0.2	$Assert(lecture, interesting)$ $\wedge\ Topic(lecture, logging)$ $\wedge\ \neg using(logging)$
0.1	JUNK

Dialogue management

- *Dialogue management* (DM) is in charge of *controlling* the conversation, and make decisions to say/do things at a given time, depending on the inputs
- Usually based on some representation of the current dialogue state

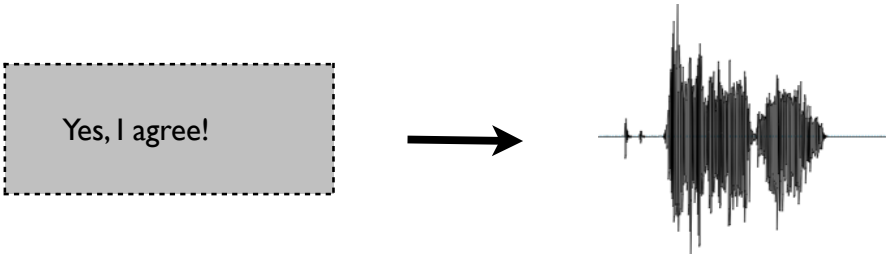
Generation

- *Natural language generation* (NLG) is the reverse task of NLU: given a high level representation of the response, *find the rights words to express it*
- How to express (or realise) the given intention might depend on various contextual factors



Speech synthesis

- Finally, speech synthesis (TTS, for «text-to-speech») is the task of generating a speech signal corresponding to the selected system response
- Can be modulated in various ways (voice, intonation, accent, etc.)



Conversation control

- Who has the initiative for the conversation?
  - *System-initiative*: the system takes the initiative (e.g. to ask questions) and the user answers
  - *User-initiative*: the user is the one controlling the interaction (he/she asks or command the machine, and the machine reacts)
  - *Mixed-initiative*: both the system and the user can take the turn at any time. Most natural, but also more complicated
- The initiative is related to the larger-question of *turn-taking* (who can talk, at what time?) which we will return to later during the course

Processing workflow

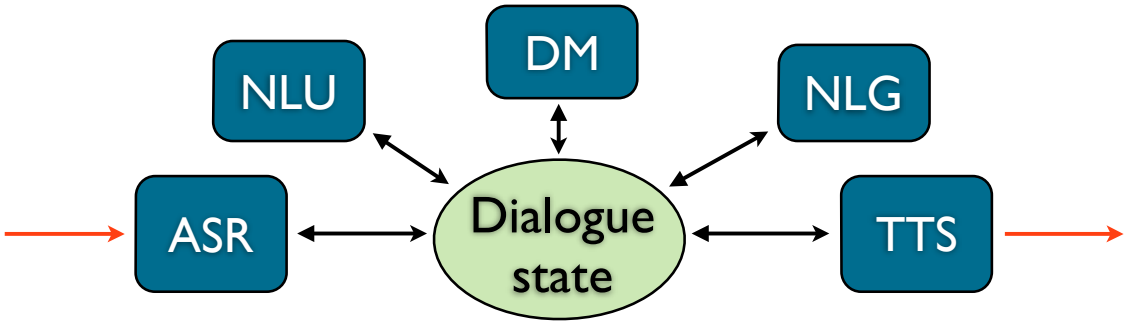
- Most basic architecture: *pipeline*
  - The components are connected in a processing chain
  - Each component is a black box getting inputs from its predecessor, and generating an output
  - Can also operate in a distributed mode

Processing workflow

- Limitations of the pipeline model
  - No feedback loop between components
  - Rigid information flow
  - Poor turn-taking behaviour (system does not react until the full pipeline has been traversed)

Processing workflow

- Blackboard (or "information-state") architectures:
  - Revolves around a *dialogue state* and a set of components
  - The modules listen for relevant changes, in which case they do some processing and update the state with the result
  - Better information flow and reactivity, but more complex design



Architectures

- Central architectural issues:
  - **Reusability**: can a specific module (e.g. a semantic parser) be reused in other dialogue systems?
  - **Domain portability**: can the system handle other dialogue domains, or is the domain «hardwired» in the system?
  - **Adaptivity**: can the dialogue system learn and adapt itself (to its user, its environment) from experience?
  - **Robustness**: can the system cope with input/output errors and module breakdowns?
  - **Efficiency**: can the system run fast enough to handle real-time interactions? Can the system run in *anytime* mode?

Outline

- Practical details
- What is a spoken dialogue system?
- Architectural schema
- **Hot research topics**
- Summary

Hot research topics

- We now review four topics of active research in spoken dialogue systems:
  - **Multimodality**: how can we build dialogue systems that can use more than one mode of communication?
  - **Situatedness**: how can we build dialogue systems that are aware of their (real or virtual) physical context?
  - **Incrementality**: how can we build dialogue systems that can process inputs and outputs as soon as possible?
  - **Statistical learning**: how can be build dialogue systems that can learn and adapt themselves from experience?

Multimodality

- **Multimodality**: The ability to communicate with the user(s) using *more than one mode of communication*
  - Can pertain both to *inputs* and *outputs*
  - *Example of multimodal input*: mobile application controlled by voice and touch
  - *Example of multimodal output*: navigation system explaining itineraries both verbally and visually on a map

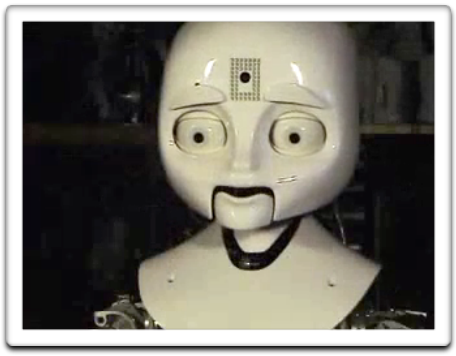
Multimodality

- For the system designer, multimodality is both a blessing and a curse
  - Using multiple input sources means more redundancy, which can help reduce the noise and uncertainty
  - Similarly, the ability to present a given *output* in several ways can improve the user experience
  - But it also adds one layer of complexity and new engineering issues, such as *synchronization* (example: a gesture combined with a spoken utterance)

Multimodality

- **Non-verbal signals** plays a crucial role in social interactions
  - Gestures, posture, affective display
- Situated dialogue systems should perceive the body language of their interlocutor (and use their own body for interacting as well)
  - Both non-verbal language *understanding* and *production*

Kismet, MIT Media Lab



Dexter, MIT Media Lab

Multimodality

- Multimodality is crucial for engagement:
  - «the process by which […] participants establish, maintain and end their perceived connection during interactions they jointly undertake»
- In many situations, the dialogue system needs to determine when a person wants to interact
  - Example: robot operating as a museum guide
- And produce the appropriate kind of verbal and non-verbal signals to keep the person engaged

[C.L. Sidner et al. (2005), «Explorations in engagement for humans and robots», *Artificial Intelligence*]

Situated context

- Many applications take place in a (real or virtual) *situated context*
- The system must take this context into account:
  - Perceive the *persons*, *objects* and *events* surrounding the agent
  - Resolve references to the external context («this person», «the blue mug», «look!»)
  - Understanding how physical actions (e.g. picking a mug) affect the state of the world

Situated context

- Artificial agents often have a very brittle understanding of their own environment
  - Detecting objects, persons, places, or events in natural environments is *really*, *really* hard
  - Imperfect sensors, unstructured and unpredictable environments, limited prior knowledge, etc..
- Situated dialogue systems must also deal with uncertainties about their perception of the surrounding context

Situated context

NB: symbol grounding ≠ dialogue grounding (feedback etc.)

- **Symbol grounding**: linguistic symbols must ultimately be grounded in other modalities
  - Bridge to *perception* and *bodily experience*
  - Linguistic symbol "door" must be linked to the prototypical image of a door
  - Role of *affordances* (what can be typically done with a door, and how)

[M. Anderson (2003), «Embodied Cognition: a Field Guide», *Artificial Intelligence*]

Incrementality

- Humans process and produce language *incrementally*
  - When listening, we don't wait for a sentence to be fully pronounced to start processing it!
  - Rather, we start our interpretation as soon as the first phonemes are available, and gradually refine our understanding as we go
  - We also continuously provide *feedback* signals
- Spontaneous human-human conversations are full of *interruptions*, *speech overlaps*, *backchannels*, and *co-completion of utterances*

Incrementality

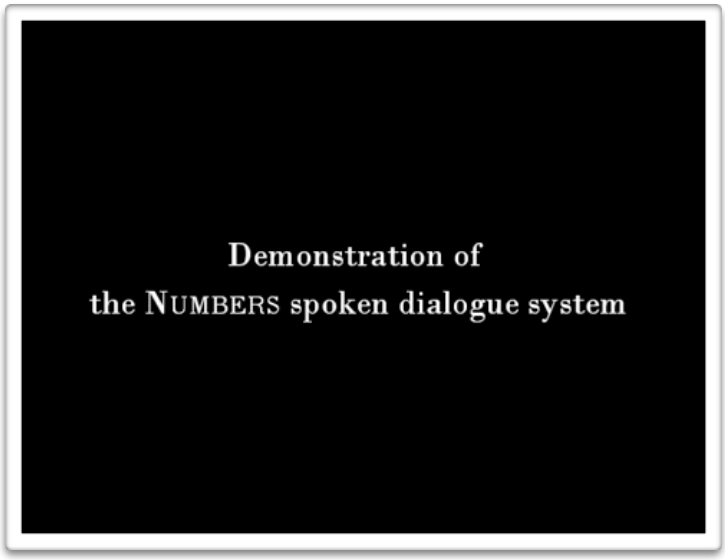
- However, most spoken dialogue systems operate in «batch mode»
  - For instance, NLU will typically wait for the sentence to be finished before parsing
  - Similarly, TTS will wait for a complete system response to start the synthesis
- This is why many dialogue systems often exhibit a «ping-pong» turn-taking behaviour:
  - Strict sequence of turns between the user and the system, one speaker at a time

Can we build **incremental** dialogue systems?

(that is, dialogue systems that can operate on "small", partial units of content)

Incrementality

Example: the NUMBERS dialogue system



[Schlangen, D. and Skantze G. (2009) «A General, Abstract Model of Incremental Dialogue Processing», in Proceedings of EACL 2009.]

Incrementality

- Advantages of incremental processing:
  - More *reactive* turn-taking behaviour
  - The system can provide feedback on its understanding (or lack thereof) while the user is speaking
  - More flexible handling of *utterance fragments*, *interruptions*
  - *Performance gains*: the system can start processing as soon as the user begins to talk

Statistical approaches

- Yet another important topic in spoken dialogue system is *statistical learning*
- Spoken dialogue systems can be hard to "hand-craft":
  - Presence of noise and uncertainty (ambiguities etc.)
  - User behaviour can be difficult to predict
  - Need to consider many alternatives

Statistical approaches

- An alternative is to rely on *machine learning* techniques (i.e. statistical modelling from data)
- Such techniques can (and have been) be applied for all components of dialogue systems:
  - Speech recognition (acoustic and language modelling)
  - Dialogue understanding (data-driven parsing etc.)
  - Dialogue management (statistical optimisation of policies)
  - Generation & speech synthesis (stochastic NLG and TTS)

[For a nice article discussing the motivation behind the use of statistics in language technology, see P. Norvig, "*On Chomsky and the Two Cultures of Statistical Learning*".]

Statistical approaches

- Advantages of statistical approaches:
  - Better account of *uncertainties* (ASR errors, ambiguities)
  - *Adaptation* of the system to its dialogue domain, external context and users
  - Principled, data-driven approach
- Current bottlenecks:
  - Good dialogue data is *scarce* and hard to acquire!
  - *Scalability* to complex domains is also an issue

Outline

- Practical details
- What is a spoken dialogue system?
- Architectural schema
- Hot research topics
- **Summary**

Summary

- Spoken dialogue systems are computational agents that can interact (bidirectionally) with humans using *spontaneous spoken language*
- They are composed of multiple components for *speech recognition* (ASR), *language understanding* (NLU), *dialogue management* (DM), *generation* (NLG) and *speech synthesis* (TTS)
- Many application domains and "open" research questions to build more intelligent, robust and adaptive dialogue systems

Next week

- Next week, we'll try to pin down in more details what spoken dialogue exactly **is**
- How are conversations structured?
- How do the participants in a conversation socially relate to each other?
- How can analyse spoken utterances?
- To answer these questions, we'll draw some insights from linguistics and cognitive science