A Salience-driven Approach to Speech Recognition for Human-Robot Interaction

Language Technology Lab
German Research Center for Artificial Intelligence (DFKI GmbH)
Saarbrücken, Germany

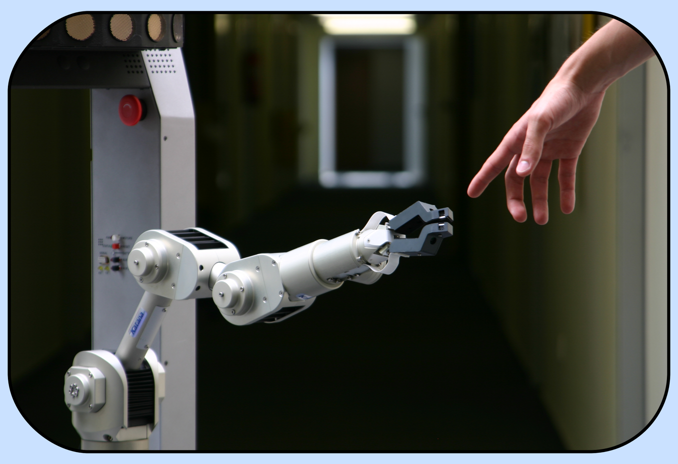
Pierre Lison
http://www.coli.uni-saarland.de/~pierrel
pierrel@coli.uni-sb.de

CoSy Project
"Cognitive systems for cognitive assistants"
EU FP6 IST
Integrated project

Background

Human-Robot Interaction

- *Interdisciplinary* research field: AI, robotics, cognitive science, comp. linguistics, and social sciences.
- Core objective: develop principles and techniques to allow efficient and natural *communication* between humans and robots
- HRI is always about *situated* interaction: language often refers to reality and discusses action and plans that affect that reality.



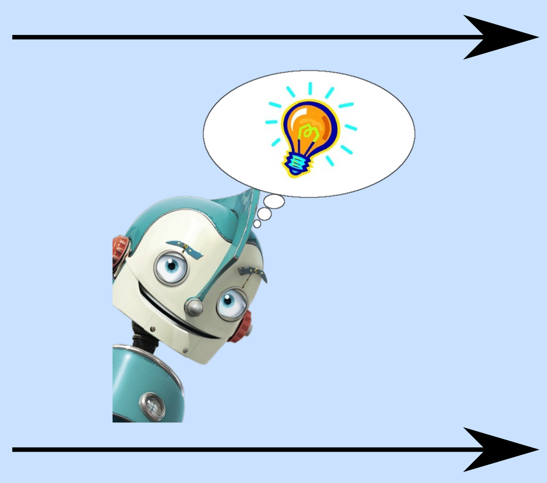
Cognitive Systems

- A cognitive system is a (artificial or biological) system able to actively *perceive* the environment it finds itself in, *reason* about it, and *achieve goals* through plans and actions.
- *Cognitive architectures* typically consist of a large number of *distributed* and *cooperating* subsystems, such as communication, computer vision, navigation & manipulation skills, and various deliberative processes (such as symbolic planners).

Key Idea

The Issue

- The first step in comprehending spoken dialogue is *automatic speech recognition* [ASR].
- The performance of speech technologies has improved significantly in the last two decades.
- But ASR remains very difficult and error-prone in the case of robots operating in real-world, noisy environments, and dealing with utterances pertaining to complex, open-ended domains.



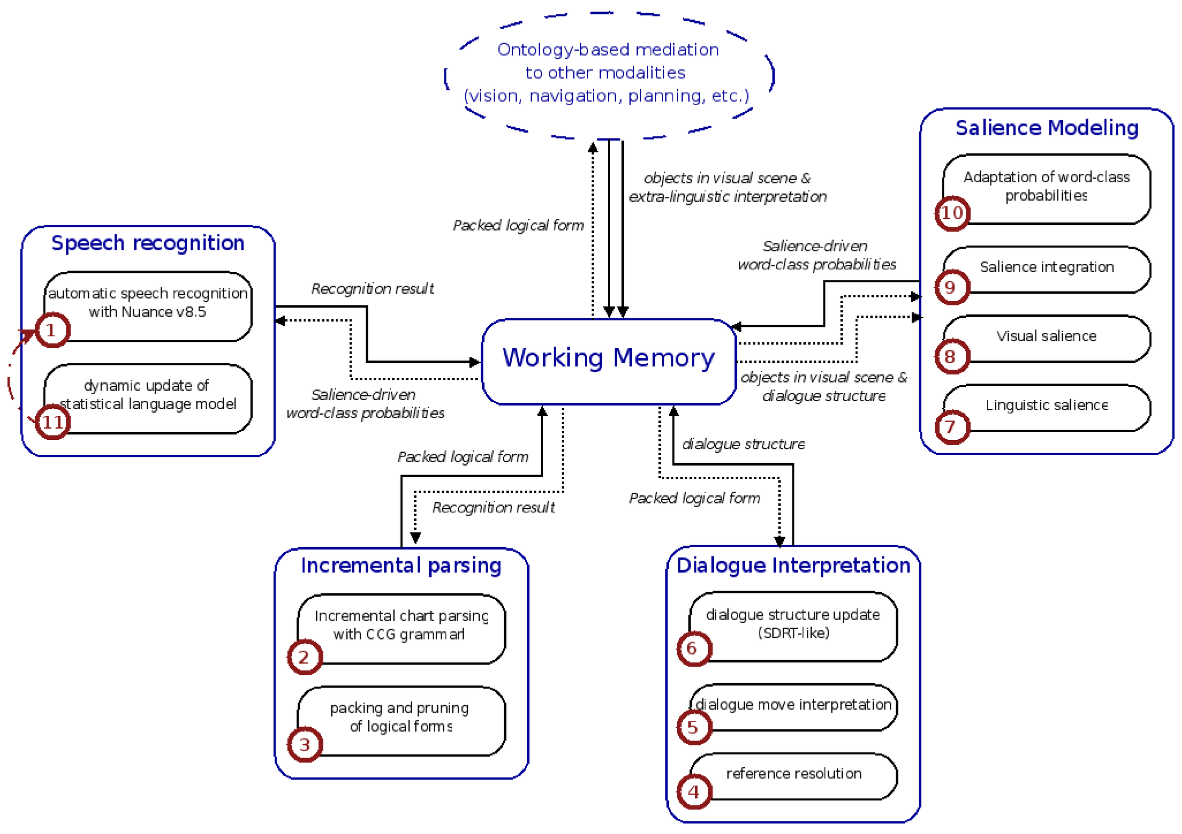
Our Hypothesis

- The intuition underlying our approach: use *contextual information* about *salient entities* in the situated environment and the dialogue state to prime utterance recognition.
- Our claim is that, in HRI, the speech recognition performance can be significantly enhanced by exploiting knowledge about the immediate physical environment and the dialogue history.

Approach

Architecture

- Our approach is implemented as part of a distributed, cognitive architecture. Each subsystem consists of several processes, and a working memory. The process can access sensors, effectors, and the working memory to access data within the subsystem.
- A specific subsystem, called the "binder", is responsible for the ontology-based mediation accross modalities
- For the ASR, we use Nuance v8.5 together with a *statistical language model* dynamically updated at runtime.



(Schematic view of the spoken dialogue comprehension module)

Salience Modeling

- *Psycholinguistic motivation*: humans systematically exploit dialogue and situated context to guide attention and help disambiguate and refine linguistic input by filtering out unlikely interpretations.
- To implement this mechanism in the robot architecture, we use *two information sources*: the salience of objects in the visual scene, and the recency of linguistic expressions in the dialogue history.
- The two saliences are integrated into a *cross-modal salience model*. We dynamically extract a set **E** of (visual and linguistic) salient entities, and compute a probability distribution P(**E**) on this set.

Language Modeling

- We define a *lexical activation network*, listing for each possible salient entity the set of its activated words. For instance, **laptop** will activate the words "laptop", "notebook", "screen, "ibm", "switch on", etc.
- The speech recognizer seeks the most likely word sequence **W*** given the acoustic input **O** and the salient objects **E**:

$$W^* = \arg\max_W \underbrace{P(O|W)}_{\text{acoustic model}} \times \underbrace{P(W|\mathbf{E})}_{\text{salience-driven language model}}$$

- For the language model, we rely on a *class-based trigram model*:

$$P(w_i|w_{i-1}w_{i-2};\mathbf{E}) = \underbrace{P(w_i|c_i;\mathbf{E})}_{\text{word-class probability}} \times \underbrace{P(c_i|c_{i-1},c_{i-2})}_{\text{class transition probability}}$$

- We introduce the salience model into the word-class probabilities:

$$P(w_i|c_i;\mathbf{E}) = \sum_{e_k \in \mathbf{E}} P(w_i|c_i;e_k) \times P(e_k)$$

- $P(w_i|c_i;e_k)$ is defined using the lexical activation: the probability of currently activated words is increased by a specific amount.

Evaluation

Controlled Experiment

- We used a test suite of 250 spoken utterances recorded during Wizard of Oz experiments. The participants were asked to interact with the robot while looking at a specific visual scene (such as the one in this box). We designed 10 different visual scenes by systematic variation of the nature, number and spatial configuration of the objects presented.
- The interactions could include descriptions, questions and commands.

Results

We focus here on the WER of our model compared to the baseline – that is, compared to a class-based trigram model not based on salience.

Word Error Rate [WER]	Classical LM	Salience-driven LM
with vocabulary = 300 words	25.48 %	24.74 %
with vocabulary = 500 words	31.74 %	27.87 %

- With a vocabulary of about 500 words, the WER is reduced by 12.2 % compared to the baseline. The difference is statistically significant (t-test).
- The salience-driven approach is especially helpful when operating with a larger vocabulary, where the expectations provided by the salience model can really make a difference in the word recognition.

References

Related to the CoSy project:
- N. Hawes et al. (2007). "Towards an Integrated Robot with Multiple Cognitive Functions". In AAAI, pp. 1548-1553. AAAI Press
- G-J. M. Kruijff, et al. (in press) "Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction", Connection Science Journal
- J. Kelleher (2005). "Integrating visual and linguistic salience for Reference Resolution". In N. Creaney (ed.), Proceedings of the 16th Irish Conference on Artificial Intelligence and Cognitive Science (AICS-05), Portstewart, Northern Ireland.

Others:
- J. Y. Chai & S. Qu (2005). "A salience driven approach to Robust Input Interpretation in Multimodal Conversational Systems". In Proceedings of HLT Conference and Conference on EMNLP 2005.
- A. Grünstein et al. (2005). "Context-sensitive statistical language modeling". In Proceedings of INTERSPEECH 2005, pp 17-20.
- K. Weilhammer, et al (2006). "Bootstrapping Language Models for Dialogue Systems". In Proceedings of INTERSPEECH 2006