Context-sensitive speech recognition for human-robot interaction

Pierre Lison & Geert-Jan M. Kruijff
Cognitive Systems @ Language Technology Lab
German Research Centre for Artificial Intelligence (DFKI GmbH)
Saarbrücken, Germany.

[ pierrel@coli.uni-sb.de ]

18th European Conference on Artificial Intelligence (ECAI 2008),
University of Patras, Greece.
July 2008.
The work we are going to present in this talk is part of the CoSy project (EU-funded integrated project).

**Goal**: “Foundations of integrated, continuously developing cognitive architectures for embodied interactive agents”.

The emphasis here will be on **interaction/communication**: we seek to develop robots which are able to

- operate in real-world, open-ended environments;
- *interact* with humans using (spoken) *natural language* to perform a variety of service-oriented tasks.
Introduction (cont’d)

Basic research question: How do we make a robot understand spoken situated dialogue?

⇒ Need to integrate a (rather sophisticated) dialogue system into the cognitive architecture.

This dialogue system must encompass multiple processing stages, from speech recognition up to semantic interpretation.

In this talk, we present a technique we developed to significantly improve the speech recognition stage.
The technique in question is a type of context-sensitive speech recognition which relies on the visual and discourse contexts to prime recognition.

The system is fully implemented as part of a cognitive architecture for robots interacting with humans using natural language to perform various service-oriented tasks.

... and has been experimentally validated on a test suite of spoken utterances, with promising results.
Outline of the talk

1. Background
   - Human-robot interaction (HRI)
   - Cognitive architectures for HRI
   - Spoken dialogue comprehension

2. Approach
   - The issue
   - Proposed solution
   - Implementation
   - Evaluation

3. Bibliography
Outline of the talk

1. Background
   - Human-robot interaction (HRI)
   - Cognitive architectures for HRI
   - Spoken dialogue comprehension

2. Approach
   - The issue
   - Proposed solution
   - Implementation
   - Evaluation

3. Bibliography
Outline of the talk

1 Background
   - Human-robot interaction (HRI)
   - Cognitive architectures for HRI
   - Spoken dialogue comprehension

2 Approach
   - The issue
   - Proposed solution
   - Implementation
   - Evaluation

3 Bibliography

Context-sensitive speech recognition for HRI
Talking robots?

Our long-term aim:

« Hi, I am C3-PO, Human Cyborg Relations. »

(And he knows over 6 million languages...)

For the time being, we’ll obviously need to scale down our expectations...
Today’s “state of the art”

Nevertheless, research in cognitive robotics is rapidly moving forward, and we are already able to do a few things:
Human-Robot Interaction

- How to make robots that actually understand what we say? And that understand why, when and how they should say something?

- Research in HRI seek to develop principles and techniques to allow efficient and natural communication between robots and humans.

- Interdisciplinary research field: artificial intelligence, robotics, (computational) linguistics, & the social sciences – psychology, cognitive science, anthropology, etc.
HRI is always about situated interaction: Language often refers to reality, discusses actions & plans affecting that reality.

Situated dialogue understanding is thus crucial for HRI.
- Understanding and producing language, relative to a current or imaginable situation in which the agents are situated.

It means that we cannot consider communication in isolation from the other modalities – we need to find meaningful ways to relate language, action and situated reality.

⇒ need to develop artificial cognitive systems able to integrate all these aspects into a common architecture [Hawes 07].
Software architectures for cognitive robots are typically composed of several *distributed* and *cooperating* subsystems, such as:

- communication;
- computer vision;
- navigation and manipulation skills;
- deliberative processes, like symbolic planners.

Our approach has been implemented as part of a *distributed cognitive architecture*. [Hawes 07].

In this talk we will focus on the subsystem for *spoken dialogue comprehension*. 
Levels of spoken dialogue comprehension

Different levels of processing:

- **Auditory**: speech recognition, (speaker localization & tracking)
- **Grammatical**: syntactic structure, semantic structure
  
  “A grammar specifies the relation between well-formed syntactic structures and their underlying (linguistic) meaning”

- **Discourse**: contextual reference resolution (anaphora, ellipsis), rhetorical relation resolution, (clarification triggers)
  
  “Discourse interprets utterance meaning relative to the established context, establishing how it contributes to furthering the discourse”
Open challenges

- **Robustness** in speech recognition:
  - noise, speaker independence, out-of-vocabulary words
  - poor performance of current ASR technology
  - (intonation, emotion)

- **Robustness** to ill-formed utterances:
  - partial, ungrammatical or extra-grammatical utterances
  - presence of various disfluencies (filled pauses, speech repairs, corrections, repetitions, etc.) in spoken dialogue.

- Pervasive **ambiguity** at all processing levels
  (lexical, syntactic, semantic, pragmatic)

- **Uncertainty** in contextual interpretation of utterances
Disfluencies in spoken dialogue: example

- Extract from a corpus of task-oriented spoken dialogue: *The Appolo Lunar Surface Journal*. [Audio file]

**Example**

**Parker**: That’s all we need. Go ahead and park on your 045. We’ll give you an update when you are done.

**Cernan**: Jack, is it worth coming right there?

**Schmitt**: Looks like a pretty good location.

**Cernan**: Okay.

**Schmitt**: We can sample the rim materials of this crater. (Pause) Bob, I’m at, let’s say, the east-southeast rim of a, oh, 30-meter crater - in the light mantle, of course - up on the Scarp and maybe 300... (correcting himself) 200 meters from the rim of Lara in a northeast direction.
Spoken dialogue comprehension

Fig.: Spoken dialogue comprehension
Spoken dialogue comprehension

**Fig.**: Spoken dialogue comprehension: step 1
Spoken dialogue comprehension

**Fig.:** Spoken dialogue comprehension: step 1
Spoken dialogue comprehension

- **Speech signal**
- **Speech Recognition**
- **Working Memory**
- **Dialogue Interpretation**
- **Mediation to other modalities**

**Fig.**: Spoken dialogue comprehension: step 1

---

Background
Approach
Bibliography

Human-robot interaction (HRI)
Cognitive architectures for HRI
Spoken dialogue comprehension
Spoken dialogue comprehension

**Fig.**: Spoken dialogue comprehension: step 2
Spoken dialogue comprehension

**Fig.**: Spoken dialogue comprehension: step 2
Spoken dialogue comprehension

**Fig.**: Spoken dialogue comprehension : step 2
Spoken dialogue comprehension

**Fig.:** Spoken dialogue comprehension: step 3

- **Speech signal**
- **Speech recognition**
- **Word lattice** to **Working Memory**
- **Incremental Parsing**
- **Mediation to other modalities**
- **Dialogue Interpretation**
Human-robot interaction (HRI)
Cognitive architectures for HRI
Spoken dialogue comprehension

Spoken dialogue comprehension

**Fig.**: Spoken dialogue comprehension: step 3
Spoken dialogue comprehension

**Fig.:** Spoken dialogue comprehension: step 3
Spoken dialogue comprehension

**Fig.**: Spoken dialogue comprehension: cross-modality
The first step in comprehending spoken dialogue is \textit{automatic speech recognition} [ASR].

For robots operating in real-world noisy environments, and dealing with utterances pertaining to complex, open-ended domains, this step is particularly error-prone.
Proposed solution

- The intuition underlying our approach: *use context!*

- More precisely, we *prime* the utterance recognition by exploiting information about
  1. The salient entities in the situated visual environment;
  2. The dialogue state.

- Our claim: for HRI, the speech recognition performance can be *significantly enhanced* by using contextual knowledge.
Psycholinguistic motivation

- Psycholinguistic studies have shown that humans do not process linguistic utterances in isolation from other modalities.

- Eye-tracking experiments notably showed that, during utterance comprehension, humans combine, in a closely time-locked fashion, linguistic information with scene understanding and world knowledge [Knoeferle 06].

- These observations, among others, provide evidence for the embodied and situated nature of language and cognition [Lakoff 87, Barsalou 99].
Practically, we use two main sources of information:

1. objects in the perceived *visual scene*;
2. linguistic expressions in the *dialogue history*.

These objects are then ranked according to their *saliency*, and integrated into a *cross-modal salience model*.

This salience model is then applied to dynamically compute *lexical activations*, which are incorporated into the language model of the speech recogniser.
Lexical activation

- A *lexical activation network* lists, for each possible salient entity, the set of words activated by it.

- In other words, it specifies *the words which are likely to be heard* when the given entity is present in the environment.

- It can therefore include words related to the object denomination, subparts, common properties or affordances.

A simple example

1. Let’s imagine we are in the lab with the robot. There is a big red ball in front of him (= high saliency).

2. The red ball is perceived by the robot sensors (camera, laser scanner, etc.), and recognised as a “red ball”.

3. In the robot’s knowledge base, the “red ball” object is associated to words like “ball” like “round”, “pick up”, etc.

4. As a final step, we adapt the language model included in the speech recogniser to increase the probability of hearing these words.
A simple example

1. Let’s imagine we are in the lab with the robot. There is a big red ball in front of him (≈ high saliency).

2. The red ball is perceived by the robot sensors (camera, laser scanner, etc.), and recognised as a “red ball”.

3. In the robot’s knowledge base, the “red ball” object is associated to words like “ball” like “round”, “pick up”, etc.

4. As a final step, we adapt the language model included in the speech recogniser to increase the probability of hearing these words.
A simple example

1. Let’s imagine we are in the lab with the robot. There is a big red ball in front of him (= high saliency).

2. The red ball is perceived by the robot sensors (camera, laser scanner, etc.), and recognised as a “red ball”.

3. In the robot’s knowledge base, the “red ball” object is associated to words like “ball” like “round”, “pick up”, etc.

4. As a final step, we adapt the language model included in the speech recogniser to increase the probability of hearing these words.
A simple example

1. Let’s imagine we are in the lab with the robot. There is a big red ball in front of him (\(\equiv\) high saliency).

2. The red ball is perceived by the robot sensors (camera, laser scanner, etc.), and recognised as a “red ball”.

3. In the robot’s knowledge base, the “red ball” object is associated to words like “ball” like “round”, “pick up”, etc.

4. As a final step, we adapt the language model included in the speech recogniser to increase the probability of hearing these words.
The objective of the situated speech recognizer is to find the word sequence $W^*$ which has the highest probability given the observed speech signal $O$ and a set $E$ of salient objects:

$$ W^* = \arg \max_W P(W|O, E) $$

$$ = \arg \max_W P(O|W) \times P(W|E) $$

For a trigram language model, the probability of the word sequence $P(w^n_1|E)$ is:

$$ P(w^n_1|E) \approx \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2}; E) $$
The objective of the situated speech recognizer is to find the word sequence $W^*$ which has the highest probability given the observed speech signal $O$ and a set $E$ of salient objects:

$$W^* = \arg \max_W P(W|O, E)$$

$$= \arg \max_W P(O|W) \times P(W|E)$$

For a trigram language model, the probability of the word sequence $P(w_1^n|E)$ is:

$$P(w_1^n|E) \approx \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2}; E)$$
Language Modeling

The objective of the situated speech recognizer is to find the word sequence $W^*$ which has the highest probability given the observed speech signal $O$ and a set $E$ of salient objects:

$$W^* = \arg \max_W P(W|O, E)$$

$$= \arg \max_W P(O|W) \times P(W|E)$$

(1)

For a trigram language model, the probability of the word sequence $P(w^n_1|E)$ is:

$$P(w^n_1|E) \approx \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2}; E)$$

(3)
Language Modeling

The objective of the situated speech recognizer is to find the word sequence $W^*$ which has the highest probability given the observed speech signal $O$ and a set $E$ of salient objects:

$$W^* = \arg\max_W P(W|O, E) \quad (1)$$

$$= \arg\max_W P(O|W) \times \underbrace{P(W|E)}_{\text{acoustic model}} \times \underbrace{P(W|E)}_{\text{salience-driven language model}} \quad (2)$$

For a trigram language model, the probability of the word sequence $P(w_1^n|E)$ is:

$$P(w_1^n|E) \simeq \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2}; E) \quad (3)$$
Our language model is class-based, so it can be further decomposed into word-class and class transition probabilities:

- Word-class probabilities: $P(w_i|c_i)$
- Class-transition probabilities: $P(c_i|c_{i-1}, c_{i-2})$.

The class transition probabilities reflect the language syntax - we assume they are independent of salient objects. The word-class probabilities, however, do depend on context.

The probability of encountering the word $w_i$ in the sequence $w_1^n$ using a class-based trigram model is therefore defined as:

$$P(w_i|w_{i-1}w_{i-2}; E) = \underbrace{P(w_i|c_i; E)}_{\text{word-class prob.}} \times \underbrace{P(c_i|c_{i-1}, c_{i-2})}_{\text{class transition prob.}} \quad (4)$$
We finally define the word-class probabilities \( P(w_i|c_i; E) \):

\[
P(w_i|c_i; E) = \sum_{e_k \in E} P(w_i|c_i; e_k) \times P(e_k) \tag{5}
\]

with \( e_k \in E \) representing a specific salient object.

To compute \( P(w_i|c_i; e_k) \), we use the **lexical activation network** specified for \( e_k \).

To put it simply, we *increase* the probability of words with are activated by \( e_k \) and *decrease* the probability of the others.

The probabilities are **dynamically updated** as the environment and the dialogue evolves and incorporated into the language model at *runtime*. 
Evaluation

- We evaluated our approach using 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.

- The evaluation results showed a significant reduction of the word error rate [WER] of our approach compared to the baseline (−16.1% for a vocabulary of about 600 words).
Thank you for your attention!!

⇒ Want to know more, and see how it works?

See my poster in the main hall!

(or check our website: www.dfki.de/cosy)
Bibliography I

L. W. Barsalou.
Perceptual symbol systems.

Towards an Integrated Robot with Multiple Cognitive Functions.
P. Knoeferle & M.C. Crocker.
*The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking.*
*Cognitive Science, 2006.*

George Lakoff.
*Women, fire and dangerous things: what categories reveal about the mind.*