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

Pierre Lison[1]

October 9, 2008

[1]Language Technology Lab,
German Research Center for Artificial Intelligence (DFKI GmbH),
Saarbrücken Germany
pierrel@coli.uni-sb.de

Abstract. We present an implemented model for speech recognition in natural environments which relies on contextual information about salient entities to prime utterance recognition. The hypothesis underlying our approach is that, in situated human-robot interaction, speech recognition performance can be significantly enhanced by exploiting knowledge about the immediate physical environment and the dialogue history. To this end, visual salience (objects perceived in the physical scene) and linguistic salience (previously referred-to objects within the current dialogue) are integrated into a single cross-modal salience model. The model is dynamically updated as the environment evolves, and is used to establish expectations about uttered words which are most likely to be heard given the context. The update is realised by continously adapting the word-class probabilities specified in the statistical language model. The present article discusses the motivations behind our approach, describes our implementation as part of a distributed, cognitive architecture for mobile robots, and reports the evaluation results on a test suite.

Keywords: human-robot interaction, speech recognition, statistical language models, salience modeling, cognitive systems.

1. Introduction

Recent years have seen increasing interest in service robots endowed with communicative capabilities. In many cases, these robots must operate in open-ended environments and interact with humans using natural language to perform a variety of service-oriented tasks. Developing cognitive systems for such robots remains a formidable challenge. Software architectures for cognitive robots are typically composed of several cooperating subsystems, such as communication, computer vision, navigation and manipulation skills, and various deliberative processes such as symbolic planners (Langley, et al. 2005).

These subsystems are highly interdependent. It is not enough to equip the robot with basic functionalities for dialogue comprehension and production to make it interact naturally in situated dialogues. We also need to find meaningful ways to relate language, action and situated reality, and enable the robot to use its perceptual experience to continuously learn and adapt itself to the environment.

The first step in comprehending spoken dialogue is *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. In spite of continuous technological advances, the performance of ASR remains for most tasks at least an order of magnitude worse than that of human listeners (Moore 2007).

One strategy for addressing this issue is to use context information to guide the speech recognition by percolating contextual constraints to the statistical language model (Gruenstein, et al. 2005). In this paper, we follow this approach by defining a *context-sensitive language model* which exploits information about salient objects in the visual scene and linguistic expressions in the dialogue history to prime recognition. To this end, a *salience model* integrating both visual and linguistic salience is used to dynamically compute lexical activations, which are incorporated into the language model at runtime.

Our approach departs from previous work on context-sensitive speech recognition by modeling salience as inherently cross-modal, instead of relying on just one particular modality such as gesture (Chai & Qu 2005), eye gaze (Qu & Chai 2007) or dialogue state (Gruenstein et al. 2005). The FUSE system described in (Roy & Mukherjee 2005) is a closely related approach, but limited to the processing of object descriptions, whereas our system was designed from the start to handle generic situated dialogues (cf. §3.3).

The structure of the paper is as follows: in the next section we briefly introduce the software architecture in which our system has been developed. We then describe the salience model, and explain how it is utilised within the language model used for ASR. We finally present the evaluation of our approach, followed by conclusions.

Figure 1: Robotic platform (left) and example of a real visual scene (right)

2. Architecture

Our approach has been implemented as part of a *distributed cognitive architecture* (Hawes, et al. 2007). Each subsystem consists of a number of processes, and a working memory. The processes can access sensors, effectors, and the working memory to share information

within the subsystem. Figure 2 illustrates the spoken dialogue comprehension. Numbers 1-11 in the figure indicate the usual sequential order for the processes..

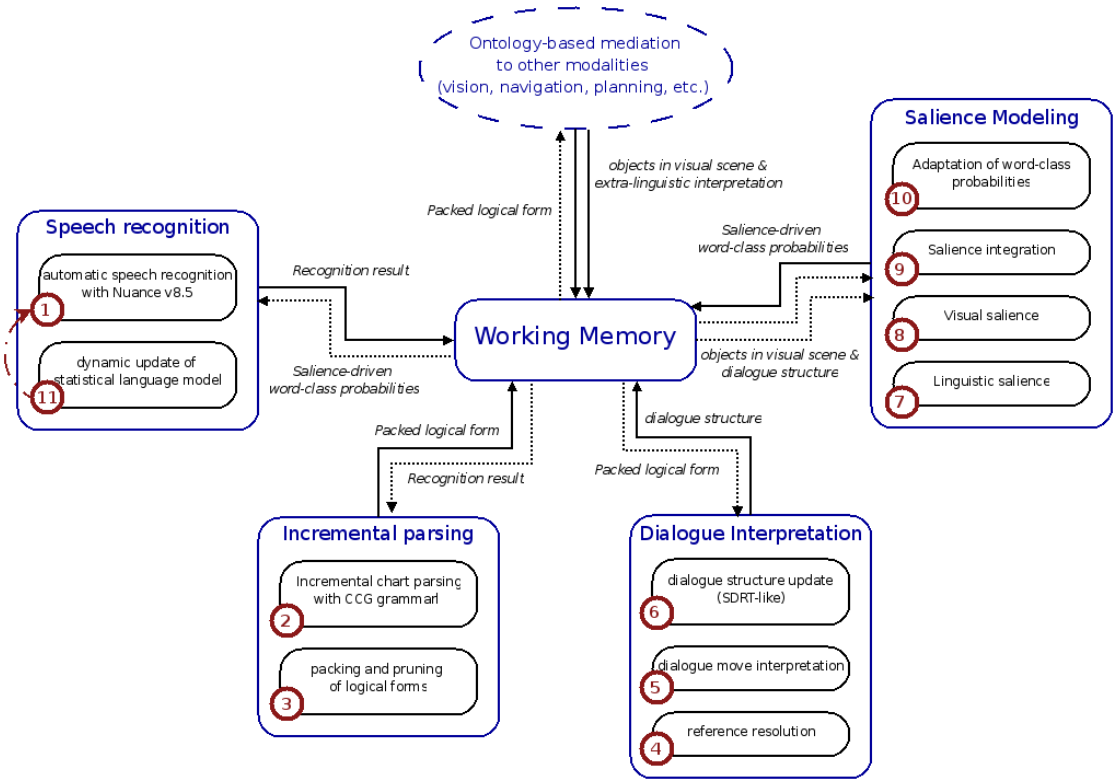

Figure 2: Schematic view of the architecture for spoken dialogue comprehension

The speech recognition utilises Nuance Recognizer v8.5 together with a statistical language model (§ 3.4). For the online update of word class probabilities according to the salience model, we use the "just-in-time grammar" functionality provided by Nuance.

Syntactic parsing is based on an incremental chart parser[1] for Combinatory Categorial Grammar (Steedman & Baldridge 2003), and yields a set of interpretations – that is, logical forms expressed as ontologically rich, relational structures (Baldridge & Kruijff 2001). Figure 3 gives an example of such logical form.

These interpretations are then *packed* into a single representation (Oepen & Carroll 2000, Kruijff, et al. in submission), a technique which enables us to efficiently handle syntactic ambiguity.

Once the packed logical form is built, it is retrieved by the dialogue recognition module, which performs dialogue-level analysis tasks such as discourse reference resolution and dialogue move interpretation, and consequently updates the dialogue structure.

Linguistic interpretations must finally be associated with extra-linguistic knowledge about the environment – dialogue comprehension hence needs to connect with other subarchitectures like vision, spatial reasoning or planning. We realise this information binding between different modalities via a specific module, called the "binder", which is re-

[1]Built on top of the OpenCCG NLP library: http://openccg.sf.net

$@_{w_1:cognition}$(**want** $\wedge$
<MOOD> ind $\wedge$
<TENSE> pres $\wedge$
<ACTOR>(i_1 : *person* $\wedge$ **I** $\wedge$
<NUMBER> sg $\wedge$
<ECOMP>(t_1 : *action-motion* $\wedge$ **take** $\wedge$
<ACTOR>y_1 : *person* $\wedge$
<PATIENT>(m_1 : *thing* $\wedge$ **mug** $\wedge$
<DELIMITATION> unique $\wedge$
<NUMBER> sg $\wedge$
<QUANTIFICATION> specific_singular)) $\wedge$
<PATIENT>(y_1 : *person* $\wedge$ **you** $\wedge$
<NUMBER> sg))

Figure 3: Logical form generated for the utterance 'I want you to take the mug'

sponsible for the ontology-based *mediation* accross modalities (Jacobsson, et al. 2008).

3. Approach

3.1. Motivation

As psycholinguistic studies have shown, humans do not process linguistic utterances in isolation from other modalities. Eye-tracking experiments notably highlighted that, during utterance comprehension, humans combine, in a closely time-locked fashion, linguistic information with scene understanding and world knowledge (Henderson & Ferreira 2004, Knoeferle & Crocker 2006).

These observations – along with many others – therefore provide solid evidence for the *embodied* and *situated* nature of language and cognition (Lakoff 1987, Barsalou 1999).

Humans thus systematically exploit dialogue and situated context to guide attention and help disambiguate and refine linguistic input by filtering out unlikely interpretations. Our approach is essentially an attempt to reproduce this mechanism in a robotic system.

3.2. Salience modeling

In our implementation, we define salience using two main sources of information:

1. the salience of objects in the perceived visual scene;
2. the linguistic salience or "recency" of linguistic expressions in the dialogue history.

In the future, other sources could be added, for instance the possible presence of gestures (Chai & Qu 2005), eye gaze tracking (Qu & Chai 2007), entities in large-scale space (Zender & Kruijff 2007), or the integration of a task model – as salience generally depends on intentionality (Landragin 2006).

3.2.1. Visual salience

Via the "binder", we can access the set of objects currently perceived in the visual scene. Each object is associated with a concept name (e.g. **printer**) and a number of features, for instance spatial coordinates or qualitative propreties like colour, shape or size.

Several features can be used to compute the salience of an object. The ones currently used in our implementation are (1) the object size and (2) its distance relative to the robot

(e.g. spatial proximity). Other features could also prove to be helpful, like the reachability of the object, or its distance from the point of visual focus – similarly to the spread of visual acuity across the human retina. To derive the visual salience value for each object, we assign a numeric value for the two variables, and then perform a weighted addition. The associated weights are determined via regression tests.

At the end of the processing, we end up with a set $\mathbf{E}_v$ of visual objects, each of which is associated with a numeric salience value $s(e_k)$, with $1 \leq k \leq |\mathbf{E}_v|$.

3.2.2. Linguistic salience

There is a vast amount of literature on the topic of linguistic salience. Roughly speaking, linguistic salience can be characterised either in terms of *hierarchical recency*, according to a tree-like model of discourse structure, or in terms of *linear recency* of mention (Kelleher 2005). Our implementation can theorically handle both types of linguistic salience, but, at the time of writing, only the linear recency is calculated.

To compute the linguistic salience, we extract a set $\mathbf{E}_l$ of potential referents from the discourse structure, and for each referent e_k we assign a salience value $s(e_k)$ equal to the distance (measured on a logarithmic scale) between its last mention and the current position in the discourse structure.

3.2.3. Cross-modal salience model

Once the visual and linguistic salience are computed, we can proceed to their integration into a cross-modal statistical model. We define the set $\mathbf{E}$ as the union of the visual and linguistic entities: $\mathbf{E} = \mathbf{E}_v \cup \mathbf{E}_l$, and devise a probability distribution $P(\mathbf{E})$ on this set:

$$P(e_k) = \frac{\delta_v \; I_{\mathbf{E}_v}(e_k) \; s_v(e_k) \;\; + \;\; \delta_l \; I_{\mathbf{E}_l}(e_k) \; s_l(e_k)}{|\mathbf{E}|} \tag{1}$$

where $I_A(x)$ is the indicator function of set A, and δ_v, δ_k are factors controlling the relative importance of each type of salience. They are determined empirically, subject to the following constraint to normalise the distribution :

$$\delta_v \sum_{e_k \in \mathbf{E}_v} s(e_k) + \delta_l \sum_{e_k \in \mathbf{E}_l} s(e_k) = |\mathbf{E}| \tag{2}$$

The statistical model $P(\mathbf{E})$ thus simply reflects the salience of each visual or linguistic entity: the more salient, the higher the probability.

3.3. Lexical activation

In order for the salience model to be of any use for speech recognition, a connection between the salient entities and their associated words in the ASR vocabulary needs to be established. To this end, we define a *lexical activation network*, which lists, for each possible salient entity, the set of words activated by it. The network specifies the words which are likely to be heard when the given entity is present in the environment or in the dialogue history. It can therefore include words related to the object denomination, subparts, common properties or affordances. The salient entity **laptop** will activate words like 'laptop', 'notebook', 'screen', 'opened', 'ibm', 'switch on/off', 'close', etc. The list is structured according to word classes, and a weight can be set on each word to modulate

the lexical activation: supposing a **laptop** is present, the word 'laptop' should receive a higher activation than, say, the word 'close', which is less situation specific.

The use of lexical activation networks is a key difference between our model and (Roy & Mukherjee 2005), which relies on a measure of "descriptive fitness" to modify the word probabilities. One advantage of our approach is the possibility to go beyond object descriptions and activate word types denoting subparts, properties or affordances of objects[2].

If the probability of specific words is increased, we need to re-normalise the probability distribution. One solution would be to decrease the probability of all non-activated words accordingly. This solution, however, suffers from a significant drawback: our vocabulary contains many context-independent words like 'thing', or 'place', whose probability should remain constant. To address this issue, we mark an explicit distinction in our vocabulary between context-dependent and context-independent words.

In the current implementation, the lexical activation network is constructed semi-manually, using a simple lexicon extraction algorithm. We start with the list of possible salient entities, which is given by

1. the set of physical objects the vision subsystem can recognise ;
2. the set of nouns specified in the CCG lexicon with 'object' as ontological type.

For each entity, we then extract its associated lexicon by matching domain-specific syntactic patterns against a corpus of dialogue transcripts.

3.4. Language modeling

We now detail the language model used for the speech recognition – a class-based trigram model enriched with contextual information provided by the salience model.

3.4.1. Corpus generation

We need a corpus to train any statistical language model. Unfortunately, no corpus of situated dialogue adapted to our task domain was available. Collecting in-domain data via Wizard of Oz experiments is a very costly and time-consuming process, so we decided to follow the approach advocated in (Weilhammer, et al. 2006) instead and generate a class-based corpus from a task grammar we had at our disposal.

Practically, we first collected a small set of WOz experiments, totalling about 800 utterances. This set is of course too small to be directly used as a corpus for language model training, but sufficient to get an intuitive idea of the kind of utterances we had to deal with.

Based on it, we designed a domain-specific context-free grammar able to cover most of the utterances. Weights were then automatically assigned to each grammar rule by parsing our initial corpus, hence leading to a small *stochastic context-free grammar*.

As a last step, this grammar is randomly traversed a large number of times, which gives us the generated corpus.

[2] In the context of a **laptop** object, 'screen' and 'switch on/off' would for instance be activated.

3.4.2. Salience-driven, class-based language models

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

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \underbrace{P(\mathbf{O}|\mathbf{W})}_{\text{acoustic model}} \times \underbrace{P(\mathbf{W}|\mathbf{E})}_{\text{salience-driven language model}} \tag{3}$$

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

$$P(w_1^n|\mathbf{E}) \simeq \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2};\mathbf{E}) \tag{4}$$

Our language model is class-based, so it can be further decomposed into word-class and class transitions probabilities. 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: for a given class – e.g. *noun* -, the probability of hearing the word 'laptop' will be higher if a laptop is present in the environment. Hence:

$$P(w_i|w_{i-1}w_{i-2};\mathbf{E}) = \underbrace{P(w_i|c_i;\mathbf{E})}_{\text{word-class probability}} \times \underbrace{P(c_i|c_{i-1},c_{i-2})}_{\text{class transition probability}} \tag{5}$$

We now define the word-class probabilities $P(w_i|c_i;\mathbf{E})$:

$$P(w_i|c_i;\mathbf{E}) = \sum_{e_k \in \mathbf{E}} P(w_i|c_i;e_k) \times P(e_k) \tag{6}$$

To compute $P(w_i|c_i;e_k)$, we use the lexical activation network specified for e_k:

$$P(w_i|c_i;e_k) = \begin{cases} P(w_i|c_i) + \alpha_1 & \text{if} \quad w_i \in \text{activatedWords}(e_k) \\ P(w_i|c_i) - \alpha_2 & \text{if} \quad w_i \notin \text{activatedWords}(e_k) \wedge \\ & \qquad w_i \in \text{contextDependentWords} \\ P(w_i|c_i) & \text{else} \end{cases} \tag{7}$$

The optimum value of α_1 is determined using regression tests. α_2 is computed relative to α_1 in order to keep the sum of all probabilities equal to 1:

$$\alpha_2 = \frac{|\text{activatedWords}|}{|\text{contextDependentWords}| - |\text{activatedWords}|} \times \alpha_1$$

These word-class probabilities are dynamically updated as the environment and the dialogue evolves and incorporated into the language model at runtime.

4. Evaluation

4.1. Evaluation procedure

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. We designed 10 different visual scenes by systematic variation of the nature, number and spatial configuration of the objects presented. Figure 4 gives an example of a visual scene.

The interactions could include descriptions, questions and commands. No particular tasks were assigned to the participants. The only constraint we imposed was that all interactions with the robot had to be related to the shared visual scene.

Figure 4: Sample visual scene including three objects: a box, a ball, and a chocolate bar.

4.2. Results

Table 1 summarises our experimental results. Due to space constraints, we focus our analysis 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**
vocabulary size $\simeq$ *200 words*	25.04 % (NBest 3: 20.72 %)	24.22 % (NBest 3: 19.97 %)
vocabulary size $\simeq$ *400 words*	26.68 % (NBest 3: 21.98 %)	**23.85** % (NBest 3: 19.97 %)
vocabulary size $\simeq$ *600 words*	28.61 % (NBest 3: 24.59 %)	23.99 % (NBest 3: 20.27 %)

Table 1: Comparative results of recognition performance

4.3. Analysis

As the results show, the use of a salience model can enhance the recognition performance in situated interactions: with a vocabulary of about 600 words, the WER is indeed reduced

by 16.1 % compared to the baseline. According to the *Sign test*, the differences for the last two tests (400 and 600 words) are statistically significant. As we could expect, 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.

The word error rate remains nevertheless quite high. This is due to several reasons. The major issue is that the words causing most recognition problems are – at least in our test suite – function words like prepositions, discourse markers, connectives, auxiliaries, etc., and not content words. Unfortunately, the use of function words is usually not context-dependent, and hence not influenced by salience. We estimated that 89 % of the recognition errors were due to function words. Moreover, our chosen test suite is constituted of "free speech" interactions, which often include lexical items or grammatical constructs outside the range of our language model.

5. Conclusion

We have presented an implemented model for speech recognition based on the concept of salience. This salience is defined via visual and linguistic cues, and is used to compute degrees of lexical activations, which are in turn applied to dynamically adapt the ASR language model to the robot's environment and dialogue state.

As future work we will examine the potential extension of our approach in three directions. First, we are investigating how to use the situated context to perform some priming of function words like prepositions or discourse markers. Second, we wish to take other information sources into account, particularly the integration of a task model, relying on data made available by the symbolic planner. And finally, we want to go beyond speech recognition, and investigate the relevance of such salience model for the development of a robust understanding system for situated dialogue.

Acknowledgements

My thanks go to G.-J. Kruijff, H. Zender, M. Wilson and N. Yampolska for their insightful comments. The research reported in this article was supported by the EU FP6 IST Cognitive Systems Integrated project *Cognitive Systems for Cognitive Assistants* "CoSy" FP6-004250-IP.

References

G. T. Altmann & Y. Kamide (2004). *Now you see it, now you don't: Mediating the mapping between language and the visual world*, pp. 347–386. Psychology Press, New York.

J. Baldridge & G.-J. M. Kruijff (2001). 'Coupling CCG and hybrid logic dependency semantics'. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 319–326, Morristown, NJ, USA. Association for Computational Linguistics.

L. W. Barsalou (1999). 'Perceptual symbol systems.'. *Behavioral & Brain Sciences* **22**(4).

J. Y. Chai & S. Qu (2005). 'A Salience Driven Approach to Robust Input Interpretation in Multimodal Conversational Systems'. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing 2005*, pp. 217–224, Vancouver, Canada. Association for Computational Linguistics.

A. Gruenstein, et al. (2005). 'Context-sensitive statistical language modeling'. In *Proceedings of INTERSPEECH 2005*, pp. 17–20.

N. Hawes, et al. (2007). 'Towards an Integrated Robot with Multiple Cognitive Functions.'. In *AAAI*, pp. 1548–1553. AAAI Press.

H. Jacobsson, et al. (2008). 'Crossmodal Content Binding in Information-Processing Architectures'. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, The Netherlands.

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.

P. Knoeferle & M. Crocker (2006). 'The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking'. *Cognitive Science* **30**(3):481–529.

G.-J. M. Kruijff, et al. (in submission). 'Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction'. *Connection Science* .

G. Lakoff (1987). *Women, fire and dangerous things: what categories reveal about the mind*. University of Chicago Press, Chicago.

F. Landragin (2006). 'Visual Perception, Language and Gesture: A Model for their Understanding in Multimodal Dialogue Systems'. *Signal Processing* **86**(12):3578–3595.

P. Langley, et al. (2005). 'Cognitive architectures: Research issues and challenges'. Tech. rep., Institute for the Study of Learning and Expertise, Palo Alto, CA.

R. K. Moore (2007). 'Spoken language processing: piecing together the puzzle'. *Speech Communication: Special Issue on Bridging the Gap Between Human and Automatic Speech Processing* **49**:418–435.

S. Oepen & J. Carroll (2000). 'Ambiguity packing in constraint-based parsing - practical results'. In *Proceedings of the 1st Conference of the North America Chapter of the Association of Computational Linguistics*, pp. 162–169, Seattle, WA.

S. Qu & J. Chai (2007). 'An Exploration of Eye Gaze in Spoken Language Processing for Multimodal Conversational Interfaces'. In *Proceedings of the Conference of the North America Chapter of the Association of Computational Linguistics*, pp. 284–291.

D. Roy & N. Mukherjee (2005). 'Towards situated speech understanding: visual context priming of language models'. *Computer Speech & Language* **19**(2):227–248.

M. Steedman & J. Baldridge (2003). 'Combinatory Categorial Grammar'. MS Draft 4.

K. Weilhammer, et al. (2006). 'Bootstrapping Language Models for Dialogue Systems'. In *Proceedings of INTERSPEECH 2006*, Pittsburgh, PA.

H. Zender & G.-J. M. Kruijff (2007). 'Towards Generating Referring Expressions in a Mobile Robot Scenario'. In *Language and Robots: Proceedings of the Symposium*, pp. 101–106, Aveiro, Portugal.