Title	Improving speech recognition on a mobile robot platform through the use of top-down visual queues
Author(s)	Ross, Robert; O'Donoghue, R. P. S.; O'Hare, G. M. P. (Greg M. P.)
Publication date	2003-08-09
Publication information	Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI-03), 9th-15th August, Acapulco, Mexico
Conference details	The 18th International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico, 9-15 August 2003
Publisher	IJCAI Workshop
Link to online version	http://ijcai.org/Past%20Proceedings/IJCAI-2003/PDF/274.pdf
Item record/more information	http://hdl.handle.net/10197/4530

Improving Speech Recognition on a Mobile Robot Platform through the use of Top-Down Visual Queues

Robert J. Ross, R.P.S. O'Donoghue and G.M.P. O'Hare

Abstract

In many real-world environments, Automatic Speech Recognition (ASR) technologies fail to provide adequate performance for applications such as human robot dialog. Despite substantial evidence that speech recognition in humans is performed in a top-down as well as bottom-up manner, ASR systems typically fail to capitalize on this, instead relying on a purely statistical, bottom up methodology. In this paper we advocate the use of a knowledge based approach to improving ASR in domains such as mobile robotics. A simple implementation is presented, which uses the visual recognition of objects in a robot's environment to increase the probability that words and sentences related to these objects will be recognized.

1 Introduction

Through the mapping of an acoustic signal to a string of words ASR systems are a key tool in the control of mobile devices such as robots, particularly where manual control is not appropriate or feasible. However despite the substantial improvements in ASR reliability which have been made in the last ten years, results can still be poor in noisy environments. Most ASR systems are based around a statistical, data driven architecture which fails to capitalize on sources of information other than the input audio signal and a static vocabulary. Hybrid architectures utilizing lip-reading algorithms have emerged in recent years [Chibelushi *et al.*, 2002], but are dependent on a user directly facing an interface.

ASR is becoming an increasingly more important tool in the development of service and mobile robots. However the speech recognition systems available fail to produce sufficient accuracy for natural interactions. Reasons for this failure include a) interference from the robot's drive systems, b) reverberations c) multiple user interference (cocktail party effect). Approaches adopted to counteract such interference range from the simplistic use of hand-held microphones, to the use of dialog systems that attempt to anticipate user utterances, allowing for the adjustment of the ASR vocabularly. [Matsui, 1999]. Although the use of microphones and dialog systems are an improvement on a purely black-box approach to ASR use, they impose tight constraints on the usability of these systems.

Although there is clear evidence that humans use both low level audio and low level visual input in the speech recognition process [McGurk and MacDonald, 1976], there is also a large body of research that indicates that humans use high level context and semantic effects to improve speech recognition [Tanenhaus *et al.*, 1995; Simpson, 1994]. Further more it is commonly observed that in conversation an individual will often reference particular themes, and discuss objects in his or her local environment.

Inspired by this evidence of context and high level semantic priming of speech recognition in humans, we propose the improvement of ASR systems on mobile robots using a top down context priming model, rather than relying on workarounds based on strict dialog systems or user held microphones. The initial model described below is based on the premise that users commonly discuss objects in their local environment. Specifically our model proposes that upon the visual recognition of an object in the environment, the probability of recognition of words related to that object is increased. This is achieved through direct communication between software agents responsible for visual processing and speech recognition. Although such a model is simplistic it acts as a stepping stone towards the further improvement of ASR by context effects.

2 Implementation

The need for a visual priming model emerged out of initial experiments with ASR performance on mobile robot platforms. We therefore present a model and implementation which have been constructed for a complete mobile robot platform, rather than having been implemented as a stand-alone algorithm.

2.1 Test Platform

Experiments are carried out using a team of Nomad Scout II robots, refitted with on-board computers, vision, and sound systems. The control system for the robot is provided through an experimental Multi-Agent System based architecture. All aspects of the robots' control, from high level planning and user modeling to low level movement and reactive control are encapsulated through a community of intentional deliberative agents.

Components relevant to the speech priming model include speech recognition, visual recognition, and semantic modeling agents. Speech recognition agents have been built around both widely used toolkits such as Sphinx and less popular hybrid systems [Carson-Berndsen and Walsh, 2000]. Visual object recognition agents employ color segmentation and edge-based feature detection algorithms to scan for objects in the robot's environment. A third important group of agents use semantic networks to provide the key filtering mechanism between the recognition of objects and the adjustment of word recognition probabilities in the speech recognition agents. A discussion of this high level speech recognition agent and the relationship between all three agent types now follows.

2.2 Speech Priming Model and Testing

A high level speech recognition agent employing a semantic network has been designed to provide probability input to lower level speech recognition agents based on the proximity of external objects. Specifically, the network has been prepared with details of objects commonly found in an office environment, along with associated verbs, prepositions and pseudonyms. When informed that there is a *letter* in the robots proximity, this agent can provide a list of words and phrases which are likely to be spoken in connection with such a letter e.g. deliver, package. This model can be altered to produce recommendations for varying periods and priorities based on relative importance of objects and the robots current actions respectively.

Typical use of this agent involves the visual recognition agent scanning the environment in search of pre-defined objects such as chairs, colored balls, and large office features. Upon detection of any such object this visual detection agent communicates its observation to any interested parties. In the context of this experiment the interested party is the high level speech recognition agent which has previously notified the visual recognition agent of its interest in any information the agent can provide. The high level speech recognition agent then notifies associated low level speech recognition agents to increase the probability of recognition of words associated with the recognized object.

Experiments conducted are based on a user and robot situated in a room which has been furnished with a number of objects, some of which the robot is capable of visually recognizing. The user then issues a number of commands to the robot. This set of commands is composed of instructions and questions about objects in the room. The command set includes both well formed commands and a number of garbled commands which are phonetically very similar to the well formed command. These incorrectly formed commands are typical of slight mispronunciations, or environmental distortions. When the context priming model is employed command recognition is clearly more reliable then that produced when the visual system is disjoint from the speech recognition components.

This method constitutes in the broadest sense a knowledge based approach to improving the speech recognition system. Such a knowledge based approach should be contrasted with what is known as multi-modal speech recognition, where the audio and visual information about the users lips are used to drive the speech recognition from the bottom up [Chibelushi *et al.*, 2002].

3 Related Work

Related research on the integration of audio and visual information includes Deb Roy's work on the learning of words from sights and sounds [Roy, 1999]. the integration of low level speech and vision information in improving speech recognition [Chibelushi *et al.*, 2002], and the generation of natural language from a visual representation [Herzog and Wazinski, 1994].

4 Initial Conclusions & Future Work

Although speech recognition systems can often produce inaccurate results in real-world environments, accuracy in the mobile robot domain can be improved through a knowledge based priming of speech systems using visual recognition sources. Such a technique is novel and should be contrasted with low level data-driven implementations. Future work includes the expansion of the semantic model, and the introduction of user and task domain modeling.

Acknowledgements

We gratefully acknowledge the support of Enterprise Ireland through grant No. IF/2001/02, SAID.

References

[Carson-Berndsen and Walsh, 2000] Julie Carson-Berndsen and Michael Walsh. Interpreting multilinear representations of speech. In *Proceeedings of the 8th Australian Conference on Speech Science and Technology*, 2000.

[Chibelushi *et al.*, 2002] C.C. Chibelushi, F. Deravi, and J.S.D. Mason. A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, 4(1):23–37, mar 2002.

[Herzog and Wazinski, 1994] G. Herzog and P. Wazinski. VIsual TRAnslator: linking perceptions and natural language descriptions. *Artificial Intelligence Review*, 8(2-3):175–187, 1994.

[Matsui, 1999] Toshihiro Matsui. Integrated natural spoken dialogue system of jijo-2 mobile robot for office services. In *AAAI/IAAI*, pages 621–627, 1999.

[McGurk and MacDonald, 1976] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, dec 1976.

[Roy, 1999] Deb Kumar Roy. *Learning Words from Sights and Sounds: A Computational Model*. PhD thesis, MIT, 1999.

[Simpson, 1994] Greg B. Simpson. Context and the processing of ambiguous words. In *Handbook of Psycholinguistics.*, chapter 10, pages 359–374. Academic Press, 1994.

[Tanenhaus *et al.*, 1995] M.K. Tanenhaus, M.J. Spivey-Knowlton, K.M. Eberhard, and J.E. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634, 1995.