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 Motivation & Proposal

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 and PDAs, particularly in cases 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 very poor in noisy environments. Most ASR systems are built around a statistical, data driven architecture which fails to capitalize on sources of information other than the input audio signal and static vocabulary. Many hybrid architectures utilizing lip-reading algorithms have emerged in recent years [Chibelushi *et al.,* 2002], but these are dependent on a user directly facing a communication interface.

ASR is becoming an increasingly more important tool in the development of service and mobile robots. This can generally be accredited to the ease of use a natural language interface should bring to human-machine interactions. However in practice the speech recognition systems available fail to produce sufficient accuracy for natural interactions. These failings result from a wide range of environmental distortions and noise, including a) interference from the robot's drive systems, b) reverberations c) multiple user interference (the so-called cocktail party effect).

Users of ASR in the robotics community have generally taken two approaches to counteracting sources of interference. On a practical level the simplistic use of hand-held or head-mounted microphones can dramatically improve performance. Another approach uses dialog systems based on finite state grammar models or frame systems to anticipate user utterances from dialog constructs [Matsui, 1999; Bischoff, 2001]. Although these approaches do offer a substantial improvement on a purely black-box approach to ASR, they impose tight constraints on the usability and flexibility of the system as a whole.

Naturally the partial failure of ASR technology in a noisy environment does not only pose problems for the mobile robotics community. Problems with mono-modal speech recognition systems have been recognized for many years, and many attempts have been made to improve recognition quality through the use of low-level visual information. In [Chibelushi *et al,* 2002] Chibelushi, Deravi and Mason present a comprehensive review of these approaches. Algorithms providing sources of visual information center mostly on those which perform lip-reading or the analysis of facial gesture. The techniques discussed in the review paper can be characterized as being low-level or data driven in nature, with a division being drawn on whether information should be combined before integration into a model, or whether two separate low-level models should first be formed before a general abstraction is made.

The trend towards integrating low level visual and audio information reflects clear evidence of this phenonimon in humans. The McGurk effect is perhaps the best known example of this, where the audio presentation of the sound 'BA' along with the visual lip movements of the sound 'GA' typically results in a listener perceiving the sound 'DA [McGurk and MacDonald, 1976]. However there is also a growing body of evidence that indicates that humans use high level context and semantic effects to improve our speech recognition performance [Tanenhaus *et al.,* 1995; Simpson, 1994]. Further more it is commonly observed that in conversation we 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 through the use of 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 associated with that object should be 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 research presented here has been conducted as part of the SAID (Speaking Autonomous Intelligent Devices) Project in University College Dublin [Ross *et al.,* 2002]. The focus of this project is to examine how the two very different computational paradigms of speech recognition and autonomous agents can benefit from each other in the production of intelligent speech enabled devices. The model developed here has therefore been constructed for a complete mobile robot platform, rather than having been implemented as a stand-alone algorithm.

Although the model presented has been developed for a mobile robot with a vision source, the basic principle is not hard-wired to the mobile robot domain. The principle can be applied anywhere where there is a source of high level information which can be used to dynamically prime a speech recognition systems. Specificly the Speech Priming agent discussed below can be applied on any mobile device which has access to spatial knowledge of the outside world (whether acquired through passive visual detection or not).

2.1 Platform

Experiments are carried out using a team of Nomaidc Scout II robots, refitted with on-board computers, vision, and sound systems. The control system for the robot is provided through MARC, an experimental Multi-Agent System based architecture [Ross, 2002]. This architecture like many social robot architectures views individual robots as intelligent agents in a social community. However MARC is original in that all aspects of any one robot's control, from high level planning and user modeling to low level movement and reactive control are encapsulated through a community of intentional deliberative agents (See [Wooldridge, 2000] for a good introduction to the theoretical models, and [Collier, 2001] for details of a concrete development and runtime environment for these agents.). Although the robot control architecture contains a large number of diverse agents, discussion here will be limited to those agents specificly connected with speech recognition with visual priming.

2.2 The Agents

The speech recognition and visual priming system consists of a number of agents which provide a) basic speech recognition, b) passive visual object recognition and c) speech recognition priming through the use of a simple semantic model. Despite the fact that these individual agents are internally highly data driven, coupling between the agents is loose with all inter-agent communication via the agent communication language Teanga [Rooney, 2001]. The internal design of the three main agent types will now be outlined, followed by a discussion of their typical interactions and usage.

Low Level Speech Recognition Agents - These agents work at the level of traditional ASR systems to convert audio signal from the outside world to simple strings of text or n-grams. A number of low level speech agents have been built using a range of speech recognition toolkits. Toolkits employed included both well-known systems like Sphinx and Via Voice, as well as a less popular but potentially more powerful and open hybrid systems [Carson-Berndsen and Walsh, 2000].

A problem encountered in using the more popular speech recognition toolkits, is that most operate using either a 13NF grammar of predicted sentences, or in a completely free-form dictation mode. A middle ground where the adjustment of probabilities on words and sentences was not achieved easily.

Visual Object Recognition Agent - Each robot has on on board color *CCD* camera. This camera is controlled by low level software and drivers to provide video footage of the outside world. A Visual object recognition agent was built which employs color segmentation and edge-based feature detection algorithms to scan for key objects in the robot's field of view. The visual recognition agent passively scans the environment for 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 agents which have previously requested to be kept informed of its findings.

Speech Recognition Priming Agent - The third agent type acts as the key filtering mechanism between the visual recognition of objects and the adjustment of word recognition probabilities in the lower-level speech recognition agents.

The agent employs a semantic network type structure to produce strings and groups of probable words for recognition based on the priority of identified external objects. Specificly, 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 package* 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 package e.g. deliver, box. The model can be altered to produce recommendations for varying periods and priorities based on relative importance of objects and the robots current actions respectively.

A typical usage scenario involves the Speech Recognition Priming Agent making an initial request with the Visual Object Recognizer for reports on any key objects recognized in the environment. If and when provided with reports of objects in the robots vicinity, the Speech Recognition Priming Agent will build up a revised list of words which it might expect a user to utter. The Speech Recognition Priming Agent will then attempt to inform the speech recognition agent of this revised list. The relevant speech recognition agent can then choose to take note of these values in speech recognition, or ignore them as it sees fit.

Based on the advice from the Speech Priming Agent, the

Speech Recognition Agents can then interprete the incoming sound signals as appropriate, providing strings of recognized text to any agents which have expressed an interest in its findings e.g. social communication agents, command interpreter agents.

2.3 Test Scenarios

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. Badly pronounced utterances range from slight mispronunciations of nouns and verbs, to the replacement of one word with a different but proper word as with

Where is the mall

rather than

Where is the ball

In initial tests the utilization of the context priming model clearly produces more reliable results than those produced when the vision system is disjoint from the speech recognition components.

3 Related Research

The integration of audio and vision is becoming more popular in recent years and is being approached from a number of angles. In addition to the low-level integration of audio and visual information as discussed earlier [Chibelushi *et al,* 2002], 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]., 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. The model presented here was based on the premise that users often discuss objects in their local environment. Through the visual recognition of objects ASR systems were 'primed' for specific topics of conversation. Such a technique is novel and should be contrasted with low level bi-modal speech recognition systems.

Although the model presented here is simple, the principle of using high level data sources to improve ASR performance can easily be extended to other data sources and platforms. Taking advantage of user and task domain modeling is an obvious next step. With this in mind, immediate future work includes the expansion of the semantic model, and the addition of general task domains.

Acknowledgements

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

References

[Bischoff, 2001] Rainer Bischoff. Hermes - a humanoid experimental robot for mobile manipulation and exploration services. IEEE International Conference on Robotics and Automation, apr 2001. Video Abstract.

[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(I):23-37, mar 2002.

[Collier, 2001] Rem W. Collier. *Agent Factory: A Framework for hte Engineering of Agent Oriented Applications.* PhD thesis, University College Dublin, 2001.

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

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

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

[Rooney, 2001] C.F.B. Rooney. A formal syntax & semantics for teanga. Technical report, University College Dublin, 2001.

[Ross *et al,* 2002] Robert J. Ross, Bryan McElency, Robert Kelly, Tarek Abu-Amer, Michael Walsh, Julie Carson-Berndsen, and Gregory M.R O'Hare. Speaking autonomous intelligent devices. In *In Proc. 13th Irish Conference on Artificial Intelligence & Cognitive Science (AICS 02),* 2002.

[Ross, 2002] Robert Ross. Marc - the multi-agent robot control architecture. Technical Report, Oct 2002.

[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.

[Wooldridge, 2000] Michael Wooldridge. *Reasoning about Rational Agents.* The MIT Press: Cambridge, MA, USA, 2000.