A Framework for Intelligent Speech Processing in Multi-Agent Environments

Nicholas Hanssens

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Metaglue, the Intelligent Room's Multi-Agent environment [4], currently handles speech input primarily in small context-free grammars which interpret local commands [3]. While this approach is functional, it has also become somewhat limiting as the sensory capabilities of the Intelligent Room grow and develop. In particular, systems which provide gesture recognition, geographic information, and awareness of states of the room (such as who is available) can all potentially add power and flexibility to the recognition engine, but are not accessible from within small, local grammars.

Motivation: The current framework for speech processing in the Intelligent Room assigns context-free grammars to individual agents. In turn, agents receive speech tags from the processor whenever a user utters a statement which corresponds to one in that agent's grammar [3]. While this approach works well for relatively simple command grammars, its locality becomes a limiting factor as the capabilities of the room grow. For example, a user may wish to turn on a light by referring to its position. While this could be encoded within a context-free grammar, it would require manually coding each light to contain grammars for position information. This procedure is highly repetitive and runs counter to the general goal of agent re-use, which would allow different instantiations of the same agent to control different lights.

Similarly, with gesture recognition becoming available in the Intelligent Room, the ability to integrate references such as "that" or "there" within spoken references is highly desirable. The strict local grammar approach is inappropriate in the case where individual arguments need to be resolved by other agents, as this would requires a high degree of repetitive, manually coded dependencies in both the grammars and the procedures. For example, if an audio multiplexier agent were to try to resolve "put that cd player on the speakers," it would require a hard-coded, local system to resolve "that," and it would need to explicitly know about every system capable of providing information about that reference. While this could be implemented, it would make the local grammars huge, massively redundant, and would require modifying the entire system every time a new agent with the ability to resolve a "that" was developed. Furthermore, the gesture recognition example fails entirely on cases such as "turn on that lamp" where the agent referenced by the command needs to be resolved by another system, as there is no way to tell which agent to pass the command to.

Additionally, rigid local grammars tend to work only with pre-set spoken commands, which makes free-form, natural interaction with the Intelligent Room difficult. A system which combined grammars and recognized the proper keys from free-form statements, a notable feature of the Galaxy system [5], could dramatically improve the ability to interact with the room naturally.

With motivations from the field of human-computer interactions, the Intelligent Room has the long-term goal of making speech interactions natural and free-form. Additionally, from the perspective of knowledge-based systems, the Intelligent Room project aims to produce computationally sound data structures which can be used in a dynamic and versatile environment. Towards these ends, this project aims to develop a solid backbone for dynamic, context-aware, intelligent speech recognition which takes full advantage of whatever sensing and processing tools are available in a given environment.

Lastly, as an Oxygen project, the integration of Spoken Language System's Galaxy [5] system for speech recognition is a high priority.

Previous Work: This project pulls heavily from previous work done in the Intelligent Room involving Metaglue [4], Metaglue's existing speech system [3], and a previous attempt to do natural speech recognition using START [7, 6].

This project also relies heavily on the work of the Spoken Language Systems group and their Galaxy software for speech recognition, largely because of its capability to move beyond context-free grammars and recognize key values as they appear in sentences [5].

Related work at other universities includes the ATT-Meta system for metaphorical reasoning at the University of Birmingham [1, 2] and work on context-based text understanding by the Text Understanding Lab at Freiburg University [8].

Approach: This project is an extension to the Metaglue system and will potentially replace the existing speech infrastructure. At the backbone of the system will be some form of SLS's Galaxy system, possibly with some components being replaced by agents within Metaglue. The Metaglue-end of the project will consist of several smaller sub-projects.

Initially, the integration of Galaxy with Metaglue and the implementation of an appropriate grammar system will provide a basis for the system. It is important that this stage be done first, as the existing context-free grammar based packages are not adequate for implementing the higher level combinations of grammars.

Following this, the next stage of the project is the development of a semantic description language, probably using XML or RDF, to describe the user-accessible commands of agents, the information-resolution capabilities of agents, and the mappings from spoken utterances to those commands. For example, a gesture recognition agent would describe its ability to resolve which object the user is pointing at and a map from that ability to words such as "this" or "that." This stage will be done in parallel with the central speech recognizer which will combine the available commands with the available information-resolution capabilities to build a grammar which will then be passed to the recognizer. Additionally, this stage may be done in conjunction with the development of a generalized meta-information architecture for describing agents within Metaglue.

Impact: This generalized architecture for registering agents not just as command recognizers, but also as information processors, adds a great degree of flexibility to the way speech recognition might be done in the future. This approach allows for speech to take advantage of whatever sensing tools are available at any given location, and to switch between tools and locations in a true dynamic fashion. Furthermore, this infrastructure allows for the rapid development of connections between sensing tools and speech processing, making possible additional research on and deployment of techniques for natural speech recognition.

Future Work: With the completion of this project, several additional optimizations will become possible. One future project is learning speech contexts (which utterances tend to be used in which isolated sets), as switching across those contexts could improve accuracy and allow the recognition grammars to be extended. Additionally, integration with resource management, person tracking, security systems, and multi-user settings will provide the basis of future speech improvements.

Research Support: This work is funded by DARPA under contract number F33615-00-C-1702, administered by AFRL / IFSC.

References:

[1] John Barnden. Combining uncertain belief resaoning and uncertain metaphor-based reasoning. In *Proceedings of Twentieth Annual Meeting of the Cognitive Science Society*, 1998.

[2] John Barnden. Uncertainty and conflict handling in the att-meta context-based system for metaphorical reasoning. In *Modeling and Using Context*, 2001.

[3] Machael Coen, Luke Weisman, Kavita Thomas, and Marion Groh. A context sensitive natural language modality for an intelligent room. In *Managing Interactions in Smart Environments*, 1999.

[4] Michael Coen. The future of human-computer interaction or how i learned to stop worrying and love my intelligent room. *IEEE Intelligent Systems*, 1999.

[5] Victor Zue et al. Jupiter: A telephone-based conversational interface for weather. *IEEE Transactions on Speech and Audio Processing*, 2000.

[6] Boris Katz. From sentence processing to information access on the world wide web. In *AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, 1997.

[7] Katherine Koch. Situating a natural language system in a multimodal environment. Master's thesis, MIT, 2001.

[8] Martin Romacker and Udo Hahn. Context-based ambiguity management for natural language processing. In *Modeling and Using Context*, 2001.