Incremental Multimodal Interface for Human Robot Interaction

Afshin Ameri E.
Mälardelan University
Högskoleplan 1, 72123,
Västerås, Sweden
aai08001@student.mdh.se

Batu Akan
Mälardelan University
Högskoleplan 1, 72123,
Västerås, Sweden
batu.akan@mdh.se

Baran Çürüklü
Mälardelan University
Högskoleplan 1, 72123,
Västerås, Sweden
baran.curuklu@mdh.se

Abstract

Face-to-face human communication is a multimodal and incremental process. An intelligent robot that operates in close relation with humans should have the ability to communicate with its human colleagues in such manner. The process of understanding and responding to multimodal inputs has been an interesting field of research and resulted in advancements in areas such as syntactic and semantic analysis, modality fusion and dialogue management. Some approaches in syntactic and semantic analysis take incremental nature of human interaction into account.

Our goal is to unify syntactic/semantic analysis, modality fusion and dialogue management processes into an incremental multimodal interaction manager. We believe that this approach will lead to a more robust system which can perform faster than today's systems.

1. Introduction

Companies producing mass market products such as car industries have been using industrial robots for machine tending, joining, and welding metal sheets for several decades. However, in small medium enterprises (SMEs) robots are not commonly found. Even though the hardware cost of industrial robots has decreased, the integration and programming costs make them unfavorable for many SMEs. In order to make industrial robots more common within the SME sector, industrial robots should easily be (re)programmable by any employee of the manufacturing plant. Our goal is to give a robot the ability to communicate with its human colleagues in the way that humans communicate with each other, therefore making the programming of industrial robots more intuitive and easy.

Speech, facial gestures, body gestures, images, etc. are different information channels that humans use in their everyday interaction; and most of the times they use more than one at the same time. At the other hand, humans see robots as objects with human-like qualities [1]. Consequently, a human-like interaction interface for robots will lead to a richer communication between humans and robots.

In-person communication between humans is a multimodal and incremental process [2]. Multimodality is believed to produce more reliable semantic meanings out of error-prone input modes, since the inputs contain complementary information which can be used to remove vague data [3]. For example, if the speech recognition system has "blue object" and "brown object" as its two best hypotheses, the wrong hypothesis can be easily ruled out if there is support from vision system that there is no blue object in the scence.

It is also accepted that some means of communication are more error-prone to special type of information than the others [3]. For example, in an industrial environment, saying "weld this to that" while pointing at the two objects, is more reliable than saying "weld the 3cm-wide 5cm-long piece to the cylinder which is close to the red cube". That's becuase the speech channel is more error-prone when it comes to defining spatial information, while visual channel is more reliable in this case.

Human brain processes different modality inputs incrementaly. This means that the processing starts as soon as the inputs start and the semantic meaning of the inputs is build up in the brain incrementally [4]. This also applies to resolving percievable context and action planning. In other words, people understand and plan their responses incrementally and they perform the action when the complete meaning of the utterances is perceived [4, 2]. This action maybe in the form of a head nod, vocal response or something that the speaker had asked them to do. In computer domain, incremental methods help to improve response times specially for verbal inputs as it uses the speech time to perform some of the required calculations [5, 6].

Our goal in this work is to implement a system which has the following charactersitics:

- Allows for integration of new modalities to the pipeline through plug-ins.
- Processes inputs as they are being perceived, incrementally.
- The incremental pipeline includes all the interaction-related subsystems such as modality fusion, action planner and dialogue manager.

Our platform is an industrial robot and we aim to make a new way of interaction between robots and

humans in industrial environments. With an early system the scenarios like this can be achieved:
- User: “pick this” [while clicking on an object]
- Robot picks up the object
- User: “put it on top of that” [while clicking on another object]

Multimodality will also allow users to utilize the system even without speaking, for example they can just choose an object and click on a location which will perform a move action from the robot to move the object to desired location.

In this work, our focus is on simultaneous multimodal interaction, which means that the information received from different modalities are complementary to each other and form a final semantic outcome containing all the information in one result.

The rest of this paper is organized as follows: in the next part, related works and motivations are discussed. In the third part we explain our methods and current state of the project. The paper ends with a brief description of plans towards completing the system and list of references.

2. Background

Since the introduction of “Media Room” in Richard A. Bolt’s paper [7], many other systems have been implemented which are based on multimodal interaction with the user. Researchers have employed different methods in implementation of such systems [14, 2, 8, 11]. All multimodal systems are common in the sense that they receive inputs from different modalities and combine the information to build a joint semantic meaning of the inputs.

Finite-state multimodal parsing has been studied by Johnston and Bangalore and they present a method to apply finite-state transducers for parsing such inputs [8]. Unification-based approaches are also studied by Johnston [9].

Fugen and Holzapfel’s research on tight coupling of speech recognition and dialogue management shows that the performance of the system can be improved if it is coupled with dialogue manager [10].

A good study on incremental natural language processing and its integration with vision system can be found here [2] and also in [15]. Incremental parsers are also studied for translation purposes [5].

Schlangen and Skantz have proposed an abstract model for incremental dialogue processing in [11].

3. Methods

The main characteristic of our approach is its incremental disposition. This means that the system will process different modality inputs as they are being received and builds up the syntactic and semantic representation of those inputs in an incremental fashion. It also means that the process will continue in higher parts of the system such as action planner and dialogue manager. These parts will start to build up a plan and a dialogue response (if needed) to incompletely perceived inputs.

Our current system involves two different modalities; but its design allows for integration of more without the need for a change in the main system. This can be achieved by developing two external components for each modality. One of the components will be responsible for gathering input information and sending it to the central parser and the other one is responsible for parsing the inputs of the specific modality upon request from the central parser. These modality-specific parsers act on a unified multimodal.

Since users tend to employ multimodal commands mostly for spatial control [12], we decided to develop a speech/mouse system as our first step for developing the multimodal interface. In this setup the user shares the view of the robot’s camera and can select objects or locations by clicking on the view while giving verbal commands. In a complex setup which contains lots of objects, such commands make a more robust system compared to a system which only acts verbally [3].

Speech recognition is performed by Microsoft Speech API 5.0 (SAPI) in command and control mode. In this mode SAPI relies on external grammar definitions in XML format. Since the grammar needs to address all the modalities, we implemented a multimodal grammar definition language. Grammars written in this format can directly be used by the parsers or can be converted to modality-specific grammars when required. The latter has been the case only for SAPI, since it is an external component and its requirements should be fulfilled.

For each modality there is a modality-specific parser which collaborates with the main parser to build up the final syntactic outcome of the inputs. Grammar language and parsers are discussed in the coming sections, followed by a brief description of our approach for modality fusion.

3.1. Grammar definition language

The grammar definition language is a modified version of Johnston’s [8, 9] unification based grammar. Our modifications give us the freedom of having as many modalities as needed and also help us to implement an easier interface for communicating with Prolog language interpreter which is used by semantic analysis and action planning systems. It also supports definition of optional and wildcard phrases.

Optional phrases are special inputs which may be perceived but are not an integral part of a sentence. Wildcards have the same definition but they differ from optional phrases in the way that we have no hint on what they may be and therefore accept any input in their

place. These are two features of SAPI which may help us in implementation of a more robust system.

Figure 1 shows a sample of the grammar language which we have named 3MG. Braces define optional phrases and colons define different modality inputs.

```
MOVE
{
      move <NP> [to] <LOCATION>;
}
NP
{
      <DET> <COLORPROP> <OBJECT>;
      [this]:SingleObjectSelect;
      [that]:SingleObjectSelect;
}
LOCATION
{
      [here]:LocationSelect;
      [there]:LocationSelect;
}
```

Figure 1. A sample of the 3MG language.

3.2. Multimodal parser

The multimodal parser is capable of receiving inputs from unlimited number of input modalities and extracts the syntactic meaning of them by using dedicated parsers for each modality. The fusion takes place at the same time with the help of multimodal grammar graph. Figure 2 shows a graphical representation of the subsystems involved.

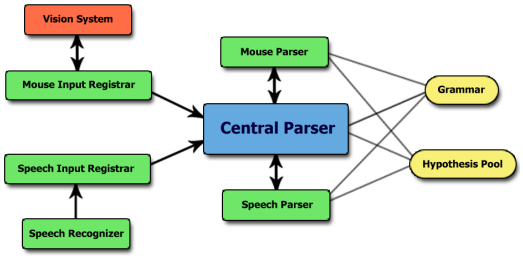

Figure 2. Multimodal parser and its subsystems in our speech/mouse setup.

The central parser constructs a grammar graph based on the 3MG grammar definition file. The graph nodes contain required data for different modalities as well as additional information regarding optional and wildcard phrases. This design allows for in-time fusion of inputs from different modalities.

Central parser also holds a pool of hypotheses. All the hypotheses in the pool have multimodal representations and therefore can be used by modality-specific parsers if needed. The central parser checks hypotheses pool after each incremental parse session and removes invalid hypotheses from the pool. It also looks for any fully parsed sentence and terminates the parse session upon finding one.

Dedicated modality parsers are modality-specific parsers which have access to the hypotheses pool. When a new parse request is received from the central parser, they try to fit the new input in the current active hypotheses and update them accordingly.

Input registrars are other modality-specific objects which are responsible for gathering input data, packing and sending it to the central parser.

Both input registrars and modality-specific parsers are designed in a way that implementing new modalities can be done with ease and without requirement of any change in the core components of the system.

3.3. Modality fusion

The central parser is also responsible for combining different modality inputs into a single final statement. The multimodal grammar graph has the key role in this process. As mentioned before, the grammar supports definition of optional phrases in it. This means that all the modality-specific parsers have the ability to skip over optional phrases if they are not perceived. The central parser takes advantage of this feature for modality fusion.

When a new input arrives in central parser, it checks the hypothesis pool to get a list of next possible phrases. If the next phrase is of the received input type the parse will continue in the modality-specific parser. But, in cases where the next phrase is of other modality than the one perceived, then the central parser makes that node as an optional node and therefore allows the modality-specific parser to continue its parser with the newly received data.

This approach allows for later integration of skipped phrases without any strict limitations on time-span and order of inputs for different modalities. Limitations only apply when the inputs belong to previous utterance or have no effect on the final outcome.

Another benefit from this design is its flexibility. Users will not have to comply with the exact definition of the grammar and the system accepts inputs if they have pauses in their speech. For example in order to give a command for moving an object to a new place, any of the following is acceptable:

- speech:”move”, click on object, click on location
- speech:”move”, click on location, click on object
- speech:”move it”, click on object, click on location

4. Early results

As the integration of the parser with the robot is not finalized, in this part we give an example of the system in our test bed, which is a 2D screen containing several rectangles.

Figure 3 shows the results when the user says “move this here” and simultaneously clicks on one of the rectangles and clicks on a location for the final position of the rectangle. Please note that the optional definition of some parts of the grammar will allow the user to use any incomplete sentences like “move”, “move this” or

even “move here” while the information from other modality is capable of filling the semantic outcome. It is also possible to use only verbal commands i.e. “move the red rectangle to the left of blue rectangle”.

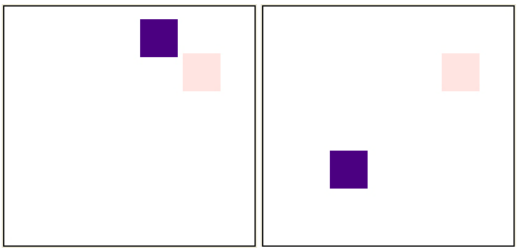

Figure 3. Multimodal test box. Before (left) and after (right) performing the move command.

This simple example can be developed to a larger grammar and used by the robot to perform different tasks based on the tool it is already using. For example, there may be commands like “drill a hole here”.

5. Future work

Since this is a work in progress, it should be noted that the project is still running and we have plans to integrate other parts of the system one by one. To read more about other works on this project you may refer to [13].

Our next step is integrating the semantic parser and vision into the incremental processing pipeline. This will give us the ability to build up semantic meaning of multimodal inputs as they are being perceived while resolving visual references that are not handled by mouse inputs.

Context manager is another component which will be added to the pipeline. Its job will be resolving in-dialogue references. The last item to integrate into the system is our dialogue manager. The incremental design of the system will make the dialogue manager able to start constructing its response as the inputs are being perceived. This may be a resolution question, a confirmation statement or a simple gesture by the robot.

References

[1] P. Schermerhorn, M. Scheutz, and C.R. Crowell, “Robot Social Presence and Gender: Do Females View Robots Differently than Males?”, Proceedings *of the third ACM IEEE international conference on human-robot interaction*, March 2008, Amsterdam, NL.

[2] T. Brick, M. Scheutz, “Incremental natural language processing for HRI”, *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, March 10-12, 2007, Arlington, Virginia, USA.

[3] S. Oviatt, “Ten myths of multimodal interaction” *Commuin. ACM,* 42(11), 74-81, Nov. 1999.

[4] Y. Kamide, G. T. M. Altmann, and S. L. Haywood. “The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements”, *Journal of Memory and Language*, 49(1), 133-156, 2003.

[5] D. Mori, S. Matsubara, and Y. Inagaki. “Incremental Parsing for Interactive Natural Language Interface”, *Proceedings of IEEE International Conference of Systems, Man and Cybernetics*, 2001.

[6] C. P. Rosé, A. Roque, and D. Bhembe, “An Efficient Incremental Architecture for Robust Interpretation”, *Proceedings of the Second International Conference on Human Language Technology Research* (March 24 - 27, 2002), Morgan Kaufmann Publishers.

[7] R. A. Bolt, “Put-that-there: Voice and Gesture at the Graphics Interface”, *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques* (July 14 - 18, 1980). SIGGRAPH '80. ACM,

[8] M. Johnston, and S. Bangalore, “Finite-state Multimodal Parsing and Understanding. *Proceedings of the 18th Conference on Computational Linguistics*, Vol.1 (July 31-August 04, 2000), Association for Computational Linguistics.

[9] M. Johnston, “Unification-based Multimodal Parsing”, *In COLING/ACL*, 1998

[10] C. Fuegen, H. Holzapfel, and A. Waibel. “Tight coupling of Speech Recognition and Dialog Management - Dialog-Context Dependent Grammar Weighting for Speech Recognition”, *Proceedings of the International Conference on Spoken Language Processing*, 2004.

[11] D.Schlangen, and G. Skantze, “A General, Abstract Model of Incremental Dialogue Processing”, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (March 30-April 03, 2009). Association for Computational Linguistics.

[12] S. Oviatt, A. DeAngeli, and K. Kuhn, “Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction”, *Proceedings of Conference on Human Factors in Computing Systems* (March 22-27), ACM Press, NY, 1997.

[13] B. Akan, B. Çuruklu, G. Spampinato, and L. Asplund, “Object Selection Using a Spatial Language for Flexible Assembly”, *Proceedings of the 14th IEEE International Conference on Emerging Technologies & Factory Automation*, (Palma de Mallorca, Spain, September 22-25, 2009). IEEE Press, Piscataway, NJ

[14] K. Hsiao, S. Vosoughi et al. “Object Schemas for Responsive Robotic Language Use”, *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction* (March 12-15, 2008). HRI '08. ACM.

[15] Geert-Jan M. Kruijff, Pierre Lison, et al. "Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction." *In Language and Robots: Proceedings from the Symposium (LangRo'2007)*. Aveiro, Portugal. December 2007.