This is to certify that the
dissertation entitled

Incorporating Non-verbal Modalities in Spoken Language
Understanding for Multimodal Conversational Systems

presented by

Shaolin Qu

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Computer Science

Major Professor's Signature

5/11/09

Date

INCORPORATING NON-VERBAL MODALITIES IN SPOKEN LANGUAGE UNDERSTANDING FOR MULTIMODAL CONVERSATIONAL SYSTEMS

By

Shaolin Qu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Computer Science

2009

ABSTRACT

INCORPORATING NON-VERBAL MODALITIES IN SPOKEN LANGUAGE UNDERSTANDING FOR MULTIMODAL CONVERSATIONAL SYSTEMS

By

Shaolin Qu

Interpreting human language is a challenging problem in building human-machine conversational systems due to the flexibility of human language behavior. This problem is further signified by insufficient speech understanding and system knowledge representation. When unreliable and unexpected language inputs are received, conversational systems tend to fail. Robust language interpretation is essential for building practical conversational systems.

To address this issue, this thesis investigates the use of non-verbal modalities for robust language interpretation in human-machine conversation. Specifically, this thesis investigates the use of deictic gesture and eye gaze to address two interrelated problems of language interpretation: unreliable speech input due to weak speech recognition, and unexpected speech input containing words that are not in the system's knowledge base. The underlying assumption is that deictic gesture and eye gaze indicate the user's visual attention and signal the salient visual context in which the user's spoken language is situated. This context constrains what the user is likely to say to the system and therefore can be used to help understand the user's language.

To facilitate this investigation, we developed a multimodal conversational system on 3D-based domains. The system supports speech, deictic gesture, and eye gaze input during human-machine conversation. Using this system, we conducted user studies to collect speech-gaze and speech-gesture data sets for the investigation. For the first topic, using non-verbal modalities to improve speech recognition and understanding, we built different salience driven language models to incorporate gesture/gaze in different stages of speech recognition. We also experimented different model-based and instance-based approaches to incorporate gesture in recognizing the intention of the user's spoken language. Our experiments show that using gesture and eye gaze significantly improves speech recognition and understanding. The use of gesture has also been shown to achieve significant improvement on user intention recognition.

For the second topic, using non-verbal modalities for automatic word acquisition, we developed different approaches to incorporate speech-gaze temporal information and domain knowledge with eye gaze to facilitate word acquisition during human-machine conversation. To further improve word acquisition, we also incorporated user interactivity to pick out the "useful" speech-gaze data for word acquisition. Our findings indicate that word acquisition is significantly improved when speech-gaze temporal information and domain knowledge are incorporated. Moreover, acquisition performance is further improved when the words are acquired from the automatically identified "useful" speech-gaze data.

The results form this thesis have important implications in building robust and practical multimodal conversational systems. They demonstrate how non-verbal modalities can be combined successfully at different stages of spoken language processing to improve robustness in language interpretation.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Joyce Chai, for her guidance and support over the years. Dr. Chai introduced me into the world of multimodal conversation and helped set up the direction of my research. Her devotion to research, commitment to professionalism, and relentless seeking of perfection have greatly inspired me through the completion of my study. I would also like to thank my guidance committee, Dr. John Deller, Dr. Anil Jain, and Dr. George Stockman for their insightful comments and suggestions that have greatly enhanced this thesis.

Many fellow graduate students have helped me for the work reported in this thesis. Special thanks to Zahar Prasov, who not only collaborated with me on the user study designs and data collection, but also had many valuable discussions with me that have helped shape my research. Tyler Baldwin, Matthew Gerber, and Chen Zhang also have contributed to the data collection and shared their valuable comments and suggestions to my work.

And finally, I want to thank my parents and my sister for their support all these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Speech is the most natural means for humans to communicate with each other. Due to its naturalness, speech is also a desirable communication mode in human-computer interaction. A lot of research has been done on spoken dialog systems [1,17,64,65,78,110], where users communicate with the system through speech. In recent years, the development of multimodal conversational systems has gained more interest. Besides speech input, multimodal conversational systems also support inputs from other modalities such as gesture and eye gaze during human-machine conversation. Compared to the conventional speech-only interfaces in spoken dialog systems, multimodal conversational interfaces provide users with greater expressive power, naturalness, and flexibility. Moreover, multimodal conversational systems can achieve better interpretation of user input due to mutual disambiguation among complementary modalities [74].

Despite recent advances in multimodal conversational systems, interpreting what a user communicates to the system is still a significant challenge due to insufficient speech recognition and language understanding performance. Moreover, when the user's utterances contain unexpected words that are out of the system's knowledge, interpretation of the user language tends to fail even when these words are correctly recognized, which also makes robust language interpretation a big challenge.

Towards building more practical multimodal conversational systems, this thesis explores the use of non-verbal modalities for robust language interpretation in two related directions. First, to improve spoken language understanding, the domain contextual information indicated by non-verbal modalities is incorporated in language modeling to get better speech hypotheses. Second, this thesis explores the use of eye gaze to acquire words automatically during human-machine conversation, in particular, by incorporating speech-gaze temporal information, domain semantic knowledge, and interactivity in word acquisition.

1.1 Overview of Multimodal Conversation

Figure 1.1 shows the typical interaction process between a user and a multimodal conversational system. The user talks to the system using speech and pen-based deictic gesture. The user's eye gaze is captured by the system. The Multimodal Interpreter identifies semantic meaning of the user's multimodal input. Given the interpretation, the Conversation Manager informs the Action Manager what action (e.g., information query, removing an object on the graphical display) to take. The Action Manager performs the action in the application domain and provides results to the Conversation Manager. Based on the results, the Conversation Manager decides what responses (e.g., inquired information not found, confirmation of object deletion) to give back to the user. The Presentation Manager presents the system's response to the user in one or more formats (e.g., audio, video, graphics).

To be able to provide intelligent responses to the user, the system first needs to understand user input, which makes Multimodal Interpreter a key component in multimodal conversational systems. This thesis focuses on building robust spoken language understanding in Multimodal Interpreter.

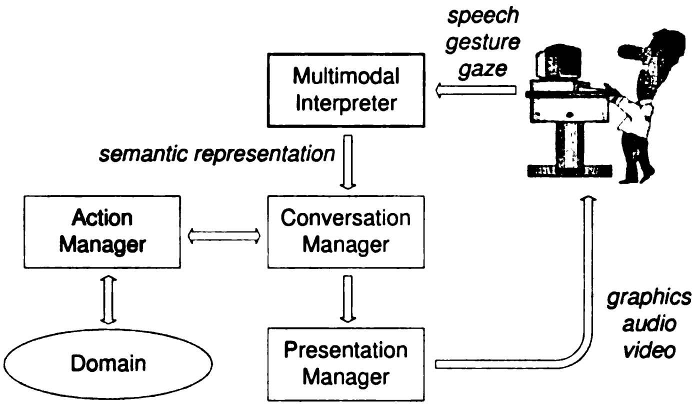


Figure 1.1. Architecture of multimodal conversation

1.2 Problems in Multimodal Language Understanding

Multimodal interpretation is to derive semantic meaning from the user's multimodal input. The interpretation process involves recognition, understanding, and integration of the user's multiple inputs of different modalities. In most multimodal conversational systems, input interpretation is based on a semantic fusion approach. In this approach, the system first creates all possible partial meaning representations independently from individual modalities. Then these partial meaning representations identified from each modality are fused in a multimodal integration process to form an overall meaning representation. Previous studies have shown that multimodal interpretation can achieve better performance than unimodal interpretation because of the mutual disambiguation among complementary modalities during multimodal integration process [74].

Figure 1.2 shows an example of the semantics-based approach to the interpretation of speech and gesture input. In the example, the user says "*what is the price of this painting?*" and at the same time points to a position on the screen. The system first creates all possible partial meaning representations from speech and gesture independently. The partial meaning representations from the speech input and the gesture input are shown in (a-b) in Figure 1.2. In this case, the gesture could be pointing to a wall or a picture. The system uses the partial meaning representations to disambiguate one and another and combines compatible partial representations together into an overall semantic representation as shown in Figure 1.2(c).

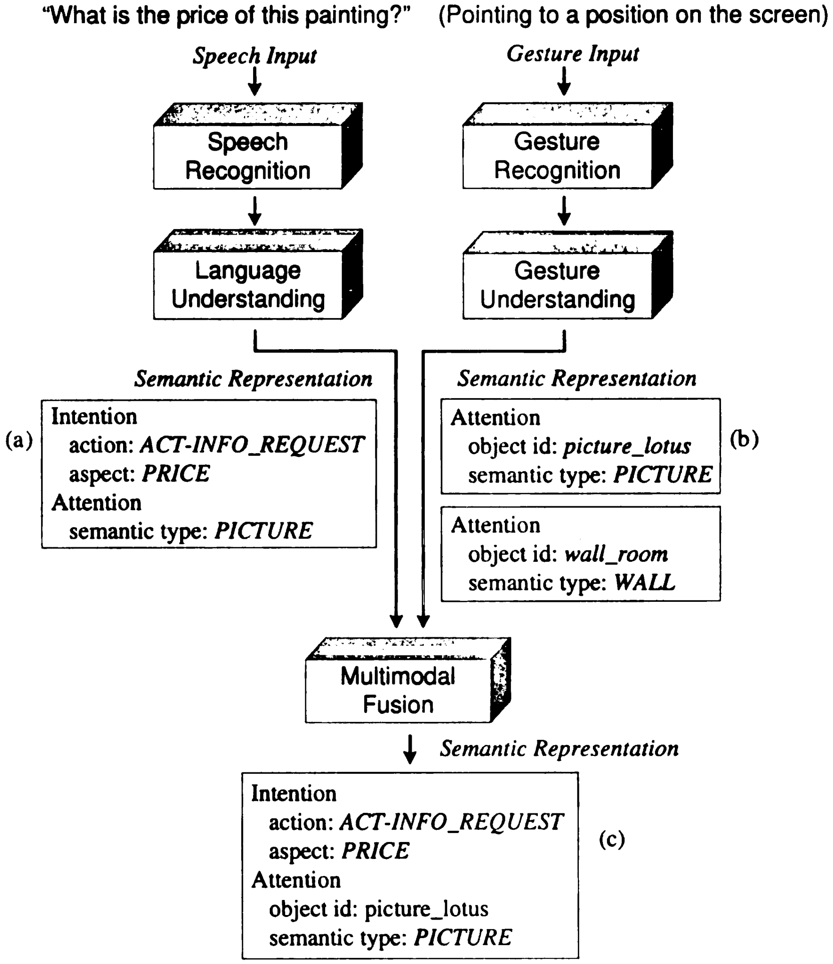


Figure 1.2. Semantics-based multimodal interpretation

In the semantics-based multimodal interpretation, the partial semantic representations from individual modalities are crucial for mutual disambiguation during multimodal fusion. A robust recognition and understanding of the user's speech is very important. However, there are two main barriers to robust spoken language understanding: unreliable speech input and unexpected speech input. We address these two problems of language understanding as follows.

1.2.1 Unreliable Speech Input

Unreliable speech input refers to the input that can not be correctly recognized due to weak speech recognition. For example, in Figure 1.2, if the speech input is recognized as "*what is the prize of this panting?*", then the partial representation from the speech input will not be correctly created in the first place. Without a correct candidate partial representation, it is not likely for multimodal fusion to reach a correct overall meaning of the input.

A potential solution to the above problem is to incorporate contextual information in recognition and understanding of speech at an earlier stage before semantic fusion in the pipelined process of multimodal interpretation. The context of human-computer interaction constrains what users are likely to interact with the system, and thus can be used to help user input interpretation. In the example, the user is talking about a picture. Suppose we already have the knowledge that the word "price" is more likely to appear in an utterance talking about a picture than the word "prize". By identifying the visual context (i.e., the picture object) from deictic gesture, the system can use the domain knowledge associated with the visual context to help recognize the word "price" correctly and thus achieve correct language understanding.

Following this idea, this thesis presents a salience driven framework in which gesture/gaze-based salience driven language models are built to improve recognized speech hypothesis. During speech recognition, these salience driven language models will guide the system to pick the speech hypothesis that is more likely describing the currently salient object as indicated by the user's gesture or eye gaze. Our experimental results have shown the potential of gesture and eye gaze in improving spoken language processing.

Besides using non-verbal modalities to obtain better speech recognition hypothesis, we also apply non-verbal modalities directly in the language understanding process to better interpret the user's spoken language, specifically, the user's intention reflected in the spoken language. In conversational systems, the "meaning" of user input can be generally categorized into *intention* and *attention* [33]. Intention indicates the user's motivation and action. Attention reflects the focus of the conversation, in other words, what has been talked about. In the speech-gesture system where speech is the dominant mode of communication, the user intention (such as asking for price of an object) is generally expressed by spoken language and attention (e.g., the specific object) is indicated by the deictic gesture on the graphical display. Based on such observations, many speech-gesture systems mainly identify intention from speech and identify attention using deictic gesture [4, 27, 53]. In our view, deictic gestures not only indicate users' attention, but also can activate the relevant domain context. This context can constrain the type of intention associated with the attention and thus provide useful information for intention recognition.

Based on this assumption, we experimented with model-based and instance-based approaches to incorporate gestural information to recognize the user's intention. We examined the effects of using gestural information for user intention recognition in two stages – speech recognition stage and language understanding stage. Our empirical results have shown that using gestural information improves intention recognition and the performance is further improved when gestures are incorporated in both speech recognition and language understanding stages compared to either stage alone.

1.2.2 Unexpected Speech Input

Unexpected speech input happens when the user speaks some words that the system can not recognize. When the encountered vocabulary is outside of the system's knowledge, conversational systems tend to fail. For example, in Figure 1.2, if the user says "*what is the cost of this painting?*" and the word "cost" is not in the system's vocabulary, then the system would not be able to understand that the user is asking for the price of the painting. Therefore, it is desirable that conversational systems can learn new words automatically during human-machine conversation. While automatic word acquisition in general is quite challenging, multimodal conversational systems offer an unique opportunity to explore word acquisition. In a multimodal conversational system where users can talk and interact with a graphical display, users' eye gaze, which occurs naturally with speech production, provides a potential channel for the system to learn new words automatically during human-machine conversation.

Psycholinguistic studies have shown that eye gaze is tightly linked to human language processing. Eye gaze is one of the reliable indicators of what a person is "thinking about" [37]. The direction of eye gaze carries information about the focus of the user's attention [49]. The perceived visual context influences spoken word recognition and mediates syntactic processing of spoken sentences [97,101]. In addition, directly before speaking a word, the eyes move to the mentioned object [31,68,88].

Motivated by these psycholinguistic findings, we investigate the use of eye gaze for automatic word acquisition in multimodal conversation. Particulary, this thesis investigates the use of temporal alignment of speech and eye gaze and domain semantic relatedness for automatic word acquisition. The speech-gaze temporal information and domain semantic information are incorporated in statistical translation models for word acquisition. Our experiment results demonstrate that eye gaze provides a potential channel for acquiring words automatically. The use of extra speech-gaze temporal information and domain semantic knowledge can significantly improve word acquisition.

Furthermore, since eye gaze could have different functions during human-machine conversation, not all speech and eye gaze data are useful for word acquisition. To further improve word acquisition, the thesis also presents approaches that automatically identify potentially "useful" speech and eye gaze based on information from multiple sources such as the user's speech, eye gaze behavior, interaction activity, and conversation context. Our experimental evaluation shows that using only the identified "useful" speech and gaze significantly improves word acquisition compared to using all speech and gaze data.

1.3 Research Questions

Addressing the above problems, this thesis investigates the following specific questions about language interpretation in speech and gesture/gaze systems:

- How can the non-verbal modalities be used to improve speech recognition?
- How can the non-verbal modalities be used to help language understanding, specifically, to help recognition of the user's intention?
- How can the non-verbal modalities be used to acquire new words automatically during multimodal conversation?

To facilitate the investigations described above, this thesis has accomplished the following objectives:

- Development of a multimodal system that supports inputs of speech, gesture and eye gaze in 3D-based domains.
- Collection of corpora of speech and gesture/gaze data from user studies.
- Design and implementation of approaches to incorporating non-verbal modalities in spoken language understanding and automatic vocabulary acquisition during multimodal conversation.
- Evaluation and analysis of these approaches that incorporate non-verbal modalities.

1.4 Road Map

The remainder of the thesis is organized as follows:

- Chapter 2: background on relevant aspects of multimodal conversation and review of previous work on multimodal language processing and language acquisition.

- Chapter 3: description of a multimodal conversational system developed for this thesis investigation. The developed system supports inputs of speech, deictic gesture, and eye gaze in a 3D interior decoration domain and a 3D treasure hunting game domain.

- Chapter 4: investigation of incorporating non-verbal modalities to improve recognized speech hypotheses for better language understanding. This chapter describes different approaches in a gesture/gaze-based salience driven framework. Evaluation and analysis of these approaches are also presented in this chapter.

- Chapter 5: investigation of incorporating non-verbal modalities to improve user intention recognition for better language understanding. This chapter describes different model-based and instance-based approaches for intention recognition and presents evaluation and analysis of these approaches.

- Chapter 6: investigation of incorporating eye gaze in automatic vocabulary acquisition for robust language understanding. This chapter describes the approaches of incorporating speech-gaze temporal information and domain semantic relatedness to facilitate word acquisition. Evaluation and analysis are also presented in this chapter.

- Chapter 7: investigation of using user interactivity related information for identifying "closely-coupled" gaze and speech streams and its effect on word acquisition. This chapter describes the prediction of "closely-coupled" gaze-speech instances for word acquisition. Evaluations of gaze-speech prediction and its effect on word acquisition are also presented in this chapter.

- Chapter 8: contributions of this thesis work.

CHAPTER 2

Background

This chapter presents a review of the topics that are relevant to this thesis. We begin by explaining the motivation for multimodal design in conversational systems, then introduce the non-verbal modalities that have been explored in multimodal conversation, and finally review the previous work on multimodal language interpretation and automatic word acquisition.

2.1 Why Multimodal Design?

One motivation for multimodal design is users' strong preference to interact multimodally. Unlike the traditional keyboard and mouse interface or a unimodal recognition-based interfaces, multimodal interfaces allow users to choose which modality to use depending on the types of information to convey, to use combined input modes, and to alternate between modes at any time. This flexible choice of input modes is preferred by users in human-computer interaction. It has been found that more than 95% percent of users chose to interact multimodally when they were free to use either speech or pen input in a map-based spatial domain [73].

Multimodal design is also motivated by the potential of multimodal systems in expanding the accessibility of computing to a broader range of users. There are large individual differences in ability and preference to using different modes of communication. These differences could be age, skill level, culture, and sensory, motor, or intellectual impairments. For example, a user with accented speech may prefer pen input rather than speech, whereas a visually impaired user may prefer speech input and text-to-speech output.

Besides expanding the range of users, multimodal systems can also expand the usage contexts. Multimodal systems allow users to switch input modes when environment condition changes or in mobile use, the user is unable to use a particular input mode temporarily. For example, users can use pen input in a noisy environment and use speech in a quiet environment, and a user of an in-vehicle multimodal application can use speech when he or she is unable to use gestural input while driving.

Another major motivation for multimodal design is the error avoidance and recovery in multimodal systems. There are user-centered and system-centered reasons why multimodal systems facilitate error recovery [75]. The user-centered reasons include:

- Users select the input mode they judge less error prone for particular lexical content, which usually leads to error avoidance. For example, in a speech and pen system, the user may prefer speech input, but will switch to pen to communicate a foreign surname.

- User's language often is simplified when interacting multimodally, which leads to better speech recognition and language understanding. For example, in a multimodal system involving a room scene, a user wants to move one of the chairs beside the bed to the window. Using only speech, the user might need to say "move the left red chair beside the bed to the window". When using both speech and gesture, the user only needs to say "move this chair here", along with two pointing gestures. This observation is most relevant to the work presented in this thesis.

- Users tend to switch modes after a system recognition error, which can prevent repeating errors and facilitate error recovery.

The system-centered reason for error recovery in multimodal systems lies in the multimodal architecture. A well designed multimodal architecture with two semantically rich input modes can support *mutual disambiguation* [74] of input signals. Mutual disambiguation involves disambiguation of signal or semantic-level information in one input mode from partial information supplied by another input mode. It leads to recovery from unimodal recognition errors within a multimodal architecture, with the net effect of suppressing errors experienced by the user. The mutual disambiguation of speech and gestural inputs has been successfully demonstrated in [14, 20, 48, 106].

2.2 Non-Verbal Modalities in Multimodal Conversational Systems

Since the appearance of Bolt's "Put That There" [4] demonstration system, which supported speech and touch-pad pointing, a variety of new multimodal conversational systems have emerged. In most of these multimodal conversational systems, the other modality besides speech is either gesture or eye gaze. Besides speech and gesture/gaze systems, there are also speech and lip movement systems where speech is processed with corresponding human lip movement information during human-computer interaction [24, 94, 102]. In speech and lip movement systems, the visual features of human lip movement is fused together with the acoustic features in the speech decoding process to perform the so-called *audio-visual speech recognition* [79].

The use of lip movement in audio-visual speech recognition is beyond the scope of this thesis. Moveover, speech recognition is not a focus of this thesis. This thesis focuses on the use of gesture and eye gaze in improving language understanding for multimodal conversation. An overview of the use of gesture and eye gaze in multimodal systems is presented as follows.

2.2.1 Gesture

In speech and gesture systems, spoken language is processed along with its accompanying gestures. The gestural input can be a simple pen-based deictic gesture (e.g., pointing, circling) [11, 15, 40, 104, 107, 108], a complex pen-based gesture involving symbolic interpretations [20,47,114], or a manual gesture [9,38,59,69].

This thesis focuses on the use of pen-based deictic gesture in spoken language processing. Deictic gesture is an active input mode, which is deployed by the user intentionally as an explicit command to the computer system. Deictic gesture has been widely used in multimodal map-based systems to indicate the focus of the user's attention (objects, locations, or areas on the map) [11,25,71,92,95,99]. Beyond only using deictic gesture as an indicator of the user's attention focus, in this thesis, we use deictic gesture to influence the recognition and understanding of the user's spoken utterances.

2.2.2 Eye Gaze

Eye gaze has been studied in various research fields such as cognitive science, psycholinguistics, and human-computer interaction. In human-computer interaction, eye gaze has long been explored for direct manipulation interfaces in which eye gaze is used as a pointing device [43,56,112,113,120]. Eye gaze as a modality in multimodal interaction goes beyond the function of pointing. In different speech and eye gaze systems, eye gaze has been explored for the purpose of mutual disambiguation [100, 121], as a complement to the speech channel for reference resolution [8,52,80] and speech recognition [21], and for managing human-computer dialogue [87]. Eye gaze has also been used as a facilitator in computer supported human-human communication [103, 105]. In this thesis, we use eye gaze and the gaze perceived visual context to help spoken language understanding in multimodal conversation.

Cognitive scientists have been studying eye movements to understand brain processes [36,88]. In psycholinguistics, eye gaze has been shown its tight link to both language comprehension [2,23,97] and language production [3,7,30]. Psycholinguistic studies have found that the gaze perceived visual context influences spoken word recognition and mediates the syntactic processing in real-time spoken language comprehension. For language production, psycholinguistic studies found that the user's eyes move to the mentioned object directly before speaking a word. These psycholinguistic findings are the motivations for this thesis's work on the use of eye gaze for spoken language processing in human-computer interaction.

Eye gaze can be captured by eye trackers, which track the user's eye movements during human-computer interaction. Two main types of eye trackers have been used in interaction study – head mounted and display mounted. Head mounted eye trackers can provide accurate gaze direction, but they are intrusive. It is unnatural and inconvenient for a user to interact with the computer system with an eye tracker mounted on the head. The state-of-the-art eye tracking technologies have enabled the eye tracking system to be embedded in a monitor. The display mounted eye trackers are non-intrusive and more appropriate for the use in human-computer interaction.

2.3 Using Non-Linguistic Information for Language Understanding

This thesis's work on using non-verbal inputs to improve spoken language understanding is inspired by previous research on multimodal language processing and context-aware language processing.

2.3.1 Multimodal Language Processing

Multimodal language processing combines speech with non-verbal modalities such as gesture, eye gaze, and lip movements for language processing. There are two levels of multimodal language processing: 1) feature-level processing; 2) semantic-level processing.

Feature-Level Processing

Feature-level processing fuses low-level feature information from parallel input signals in a multimodal architecture. Feature-level processing is most appropriate for closely synchronized modalities such as speech and lip movements. In audio-visual speech recognition [79], features of speech and lip movements are first extracted by acoustic signal processing and vision analysis respectively. The extracted audio features and visual features are then fused together for speech decoding.

Feature-level multimodal integration of speech and lip movement is beyond the scope of this thesis. This thesis investigates the use of deictic gesture and eye gaze in multimodal language processing. These modalities do not have the close coupling with acoustic speech as lip movement does, so the feature-level processing is not appropriate. Moreover, this thesis focuses on language understanding rather than speech recognition. In audio-visual speech recognition, extracted acoustic and visual features are fused for speech decoding. In this thesis, gesture/gaze is incorporated in language modeling to tailor speech hypotheses for better semantic interpretation.

Semantic-Level Processing

Semantic-level processing is to integrate semantic information derived from parallel input modes in a pipelined multimodal architecture (as seen in Figure 1.2). Semantic-level processing is mostly used for less coupled modalities such as speech and gesture. In semantic-level processing, the system first recognizes each modality independently and then creates all possible partial semantic representations individually for each modality. Then the system uses these partial semantic representations to disambiguate each other and form a joint semantic representation [10, 44, 45]. This fusion of multimodal input at the semantic level is called *late fusion* [76].

Late semantic integration systems use individual recognizers for different input modes. These individual recognizer can be trained using unimodal data, which are easier to get and already publicly available for modalities such as speech [18] and handwriting [41, 61]. Multimodal systems based on semantic fusion can also take advantage of the existing relative mature unimodal recognition techniques and off-the-shelf recognizers, which can be directly integrated in the late semantic integration architecture. In this respect, multimodal systems based on semantic fusion can be scaled up easier in the number of input modes.

Previous work on semantic fusion of multimodal input has been more focused on the integration of speech and gesture, especially pen-based gesture, than on integration of speech and eye gaze. In multimodal interaction, pen-based gesture is a much more reliable input mode for object selection than eye gaze. Moreover, pen-based gesture can contain more semantic meaning by drawing symbols or writing letters.

Due to the limitation of eye gaze, multimodal integration of speech and eye gaze has mainly been studied for simple object selection and reference resolution. In the experiments of object selection [100, 121], the user selects an object (icon) on the screen using speech, user's speech and eye gaze are both used to decide the selected object by the system. In [121], both speech and eye gaze of user generate an n-best list of potential objects, the system decides the selected object by taking the common one on both n-best lists. In [100], the selected object is decided by computing the posterior probabilities of the objects on screen being selected by the multimodal input. In the applications of reference resolution [8, 52], the object that is fixated by eye gaze prior to user's mention of the object in speech is taken as the referent for simple commands like "move it there" and "open the door".

Integration of speech and gesture for multimodal interpretation is more mature than integration of speech and eye gaze. Many integration approaches have been explored for a variety of speech and pen-based gesture systems. Those integration approaches can be categorized into the following types based on their integration mechanisms: *frame-based approaches*, *unification-based approaches*, *finite-state approaches*, *optimization-based approaches*, and *statistics-based approaches*.

Frame is a data structure used for knowledge representation. A frame has a number of slots in it. The slots represent object properties, actions, or an object's relation with other frames. Frame-based multimodal integration approaches use individual frames to represent semantic meanings obtained from different modalities and achieve multimodal integration by merging those complementary individual frames to one unified frame. Frame-based integration approaches have been used in speech and gesture systems for applications such as multimodal text editing [109], multimodal drawing [93], and multimodal appointment scheduling [106]. Frame-based approaches are simple and efficient, but they are specific to application.

Unification-based approaches are derived from computational linguistics, in which formal logics of *typed feature structures* have been well developed. The primary operation in the logics of feature structure is *unification* – determining the consistency of two feature structures and combining them into a single feature structure if they are consistent. Using feature structures for meaning representation, unification-based approaches achieve multimodal integration by performing *unification* operation over the feature structures of different modalities. Compared to frame merging, unification of typed feature structures provides a more general, formally well-understood, and reusable mechanism for multimodal integration. Unification-based approaches haven been used in the QuickSet system for the integration of speech and pen-based gesture input [44,48].

Johnston and Bangalore [45,46] employed finite-state transducers to achieve multimodal integration in a multimodal messaging application, in which users interact with a company directory using synergistic combinations of speech and pen input.

Multimodal context-free grammar (CFG) was introduced for integrating speech and gesture with finite-state transducers. The finite-state approach enables a tighter coupling of speech and gesture by using gesture to guide speech recognition, which can lead to improved speech recognition and understanding. However, the finite-state approach has one major limitation in that it requires a multimodal grammar to be created to define the language allowed in a particular application domain, which makes them only applicable for very constrained domains that involves small vocabulary and simple expressions.

Optimization-based approaches use optimization methods of machine learning for multimodal integration. Chai et al. [14] modeled integration of multimodal inputs as graph matching and applied the graph-based approach to achieve reference resolution in a map-based real estate domain, where users use speech and gesture to inquire estate information. In [27], for the purpose of multimodal reference resolution, gestures and spoken words are aligned by minimizing a penalty function defined to penalize the gesture-speech bindings that violate the empirically preferred binding rules.

Wu et al. [116] proposed a statistical hierarchical framework, Members-Teams-Committee (MTC), for the integration of speech and gesture in a simulated community fire and flood control domain. In this framework, all possible multimodal interpretations are predefined and the interpretation of a multimodal input is decided by the posterior probabilities of unimodal speech and gesture recognition hypotheses and the statistics of predefined multimodal interpretations. Since this statistics-based approach requires all possible speech and gesture interpretations to be pre-defined for a particular domain, it is only appropriate for constrained domains involving simple speech and gesture commands.

In the above late semantic fusion approaches, information from multiple modalities is only used at the fusion stage. Some low probability information (e.g., recognized alternatives with low probabilities) that could turn out to be very crucial in terms of the overall interpretation may never reach the fusion stage. Therefore, it is desirable to use information from multiple sources at an earlier stage, for example, using one modality to facilitate semantic processing of another modality. Addressing this problem in late semantic fusion, Chapter 4 of this thesis presents the use of deictic gesture and eye gaze in an earlier stage to facilitate language processing before semantic fusion.

2.3.2 Context-aware Language Processing

The context of human-computer interaction constrains what a user is likely to interact with the system, thus can be utilized for user language interpretation. A variety of research work has been done on using contextual information for spoken language processing. There are mainly two types of context used in context-aware language processing: *conversation context* and *domain context.*

All information related to the discourse prior to an utterance constitutes the conversation context of the utterance. Chotimongkol and Rudnicky [16] used conversation contextual feature to improve speech recognition and understanding by rescoring the n-best output of speech recognizer with a linear regression model. The conversation contextual feature was represented by the correlation of the current user utterance and the previous system utterance. Solsona et al. [96] combined conversation context-specific finite state grammar (FSG) and general n-gram model to improve speech recognition for a conversational system. The conversation context was represented by the types of previous system prompts and questions. Lemon and Gruenstein [62] also built conversation context-specific grammars to improve speech recognition and understanding. The conversation context was represented by the types of dialog move. Gruenstein et al. [34] built context-sensitive class-based n-gram model to improve speech recognition for a flight reservation system. The conversation context was represented by the current *information state*, which indicates whether certain information about the flight has been collected from previous conversation.

All domain related information constitutes the domain context, which could be the visual content of the graphical display in a domain, or the task knowledge in a specific domain application. Roy and Mukherjee [89] incorporated visual domain context in language model to improve spoken language comprehension in a synthetic visual scene description domain. The visual context was represented by the visual features (e.g., color, size, shape) of the objects in the scene. Coen et al. [19] built visual context-specific grammars to improve speech recognition and understanding in an Intelligent Room where a user can operate computer controlled devices by speaking. What is currently nearby the user in the room constitutes the visual context. Carbini et al. [9] used domain contextual information to help interpretation of ambiguous speech-gesture commands and enable short multimodal commands in a chess game domain. The domain contextual constraints include the displacement rules of chess game and current game position. Gorniak and Roy [29] incorporated both physical domain context and conceptual domain task related context to resolve spoken referring expressions in a 3D game domain. The physical context includes information about the physical objects in the game, such as location and type of the objects. The conceptual context consists of a set of hierarchical plan fragments to complete the specific task of the game. Due to the constrained game setting, users must follow certain steps to complete the task. Therefore, given the previous steps (physical context) and the hierarchical plan segments (task conceptual context), it is possible to predict which plan fragment the user will take, specifically which object the user will likely to refer to in his/her spoken commands.

Motivated by context-aware language processing, Chapters 4 & 5 of this thesis investigate the use of domain contextual information for improving speech recognition and understanding. Different from the context in previous work, the domain context in this thesis work is dynamically signaled by non-verbal modalities such as gesture and eye gaze during multimodal conversation. Cooke [21] also explored the use of eye gaze for spoken language processing in a map route description domain. In [21], eye gaze was used to improve speech recognition by rescoring the n-best list of speech recognition with the landmark-specific n-gram models that correspond to the gaze fixated landmarks. Different from [21], in this thesis, we explore more ways of integrating eye gaze in spoken language processing and present a better integration strategy than n-best list rescoring for the use of eye gaze in speech recognition.

2.4 Automatic Word Acquisition

Word acquisition is to learn the semantic meanings of new words. In this thesis, we focus on the automatic word acquisition by a computer system during human-computer interaction. The purpose of automatic word acquisition is to enlarge the system's knowledge base of vocabulary and therefore better interpret the user's spoken language.

In the conversational systems with which users interact through a visual scene, users talk to the system based on what is being shown on the scene and the system "understands" the user's language by mapping the spoken words to the semantic concepts in its domain knowledge base. These semantic concepts of words represent the visual entities and their properties in the domain. For these systems, the specific task of word acquisition is to ground words to the visual entities and their related properties in the domain. Word acquisition by grounding words to visual entities has been studied in various language acquisition systems.

Sankar and Grorin [91] acquired words by grounding words to visual properties (color, shape) of objects in a synthetic blocks world, in which the user interacted with the system by typing sentences. The system started with no semantic associations of words and visual properties. The only innate knowledge of the system was the semantic level signal "good" and "no". During the human-computer interaction, the user instructed the system to focus on certain objects and gave responses (e.g., "good", "no") indicating whether the system followed the instructions correctly. The goal of the system was to learn to focus on the object that the user referred to by building associations of words and visual properties. The mutual information between the occurrences of words and object shape/color types was used to evaluate the strength of the association of a word and a color/shape type.

Roy and Pentland [90] proposed a computational model that could learn words directly from raw multimodal sensory input. In their experiments, infant caregivers were asked to play toys with their infants while giving infant-directed speech. Given speech paired with video images of single objects (toys), the temporal correlation of speech and vision was used to learn words by associating the automatically segmented acoustic phone sequences with the visual prototypes (color, shape, size) of the objects.

Yu and Ballard [118] investigated word learning in a visual scene description domain in which users were asked to describe nine office objects on a desk and how to use these office tools. Given speech and the co-occurring video images captured by a head-mounted camera, a generative model was used to find the associations of automatically recognized spoken words and visual objects.

Towards the goal of robust multimodal interpretation, this thesis explores the use of eye gaze for automatic word acquisition. Eye gaze is an implicit and subconscious input, which brings additional challenges into word acquisition. Eye gaze has been explored for word acquisition in [117], in which eye gaze and other non-verbal modalities such as the user's perspective video image and hand movement were used together with speech to learn words. In the experiments, users were asked to describe what they were doing while performing three required activities: "stapling a letter", "pouring water", and "unscrewing a jar". Head-mounted eye tracker and camera were used to capture gaze and video data. Given speech paired with gaze positions and video images, a translation model was used to associate acoustic phone sequences to the four objects and nine actions in the domain.

Liu et al. [66] also investigated the use of eye gaze for word acquisition. In [66], speech and eye gaze data were collected from simplified human-computer conversation in which users verbally answered the system's questions about the decoration of a 3D room. A translation model was used to acquire words from transcribed speech and its accompanying gaze fixations.

This thesis's work on the use of eye gaze for word acquisition is different from previous work. Besides gaze positions, we use extra information such as speech-gaze temporal information and domain semantic knowledge to facilitate word acquisition. Moreover, not all co-occurring speech and gaze data are useful for word acquisition. This was not considered in the previous work on using eye gaze for word acquisition. In this thesis, we investigate the automatic identification of "useful" speech and gaze fixations and its application on word acquisition.

CHAPTER 3

A Multimodal Conversational System

To explore the incorporation of non-verbal modalities in language interpretation during multimodal conversation, we built a multimodal conversational system that supports speech, deictic gesture, and eye gaze inputs. This chapter presents the architecture of the system and the processing of different input modalities.

3.1 System Architecture

Our multimodal conversational system is built on a client/server architecture as shown in Figure 3.1. In this architecture, the user interacts with the client, a graphic interface, using speech and other modality (e.g., deictic gesture, eye gaze). The results of speech recognition and gesture/gaze recognition are sent to the server via TCP/IP network. The Multimodal Interpreter derives semantic meaning of the user's multimodal input and sends the interpretation result to a dialog manager. The Dialog Manager controls the interaction flow and decides what the system should do based on the interpretation of the user's input. The Presentation Manager decides how to present the system's responses to the user and transmits the responses to the client through the network. The system's responses are presented to the user on the client by graphics or/and speech.

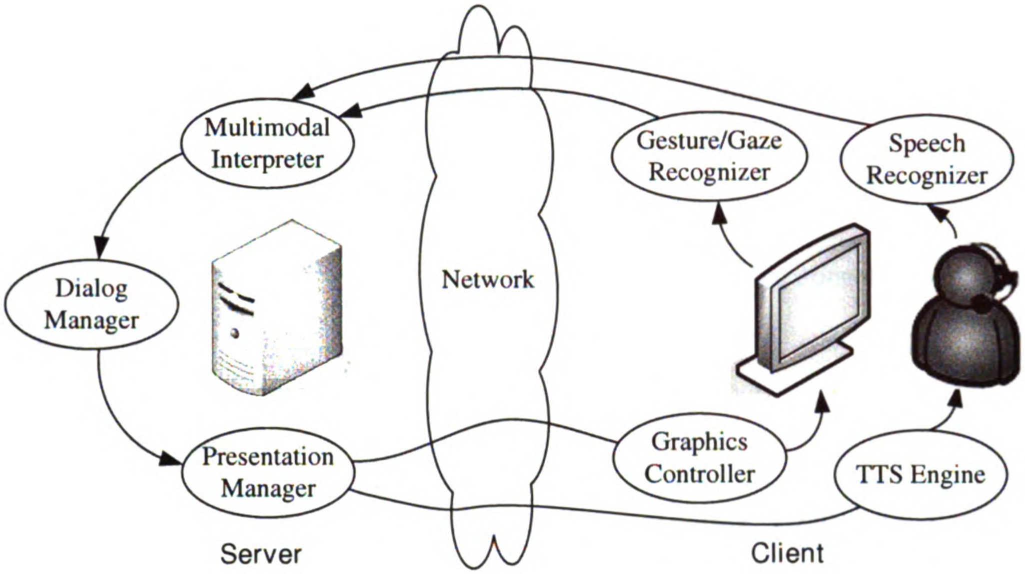


Figure 3.1. Multimodal conversational system architecture

3.2 Input Modalities

Users can interact with our multimodal conversational system using speech, deictic gesture, and eye gaze.

3.2.1 Speech

As the major input mode in multimodal conversational systems, speech enables users to interact with the system naturally and efficiently. To be able to give intelligent replies to the user, the system first needs to recognize the user's speech. Speech recognition is to convert acoustic speech signals to text. Automatic speech recognition (ASR) has been progressing steadily in the last three decades, which have resulted in commercial ASR systems that can recognize human speech with sufficient accuracy under optimal conditions. However, during natural conversation, environment noise and disfluency in users' speech can deteriorate speech recognition performance significantly. Accents in users' speech can also make speech recognition difficult. Because of these reasons, speech recognition remains a major bottleneck for building robust conversational systems.

The CMU Sphinx-4 speech recognizer [111] is used in our system for recognizing users' spoken utterances. Sphinx-4 is an open source speech recognizer based on Hidden Markov Model (HMM).

How non-verbal modalities can be incorporated to improve speech recognition is presented in Chapter 4.

3.2.2 Deictic Gesture

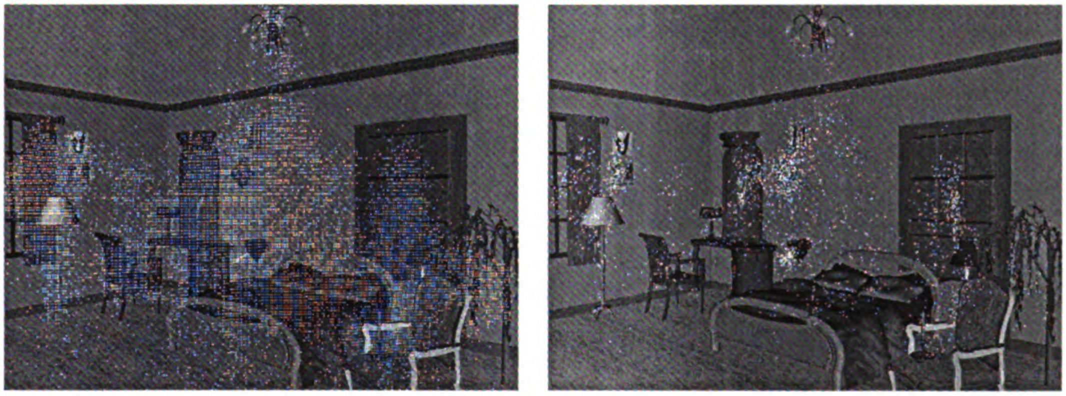
Besides speech, users can use deictic gesture (e.g., pointing, circling on a graphical display) to make interaction easier. For example, instead of say "*how much is the red chair in the left corner?*", the user can say "*how much is this chair?*" while pointing to the attended chair on the screen.

In our system, users' deictic gestures are captured by a touch screen. Based on the position of the gesture on the screen, we can infer which object the user is referring to. How this gestural information can help recognize and understand the users' speech is presented in Chapter 4 and Chapter 5.

3.2.3 Eye Gaze

Eye gaze indicates the user's focus of attention [26, 49, 101]. The published results on eye gaze and human language production have led to the hypothesis that users tend to look at the objects on the graphical display when they are talking about them. Based on this hypothesis, by tracking the user's eye gaze during human-machine conversation, the system is likely to infer the user's attended objects on the screen and use this attention information to help recognize and understand the user's speech. Moreover, using eye gaze information, the system can potentially learn new words from the user's language by associating semantics of the attended objects (indicated by eye gaze) with words in the user's spoken utterances.

Eye gaze is captured by an eye tracker. The raw gaze data points consists of the screen coordinates of each gaze point with a particular timestamp. As shown in Figure 3.2(a), this raw data is not very useful for identifying fixated objects. The raw gaze data is processed to eliminate the invalid and saccadic gaze points, leaving only pertinent eye fixations. Invalid gaze points occur when users look off the screen. Saccadic gaze points occur during ballistic eye movements between fixations. Vision studies have shown that no visual processing occurs in the human mind during saccades (i.e., saccadic suppression). It is well known that eyes do not stay still, but rather make small, frequent jerky movements. In order to best determine fixation locations, nearby gaze points are averaged together to identify fixations. The processed eye gaze fixations is shown in Figure 3.2(b).



(a) Raw gaze points

(b) Processed gaze fixations

Figure 3.2. Eye gaze on a scene

How eye gaze information can be used in language models to potentially help spoken language processing is presented in Chapter 4. How eye gaze information is used for automatic vocabulary acquisition in multimodal conversation is presented in Chapter 6.

3.3 Domains of Application

Two application domains were designed and implemented for our investigation. Both domains were constructed based on 3D graphics.

3.3.1 Interior Decoration

Figure 3.3 shows the 3D interior decoration domain. In this domain, users can interact with the system using both speech and deictic gestures to query information about the entities or arrange the room by adding, removing, moving, and coloring the entities. For example, the user may say "*remove this lamp*" or ask "*what's the power of this lamp?*" while pointing at a lamp in the scene.

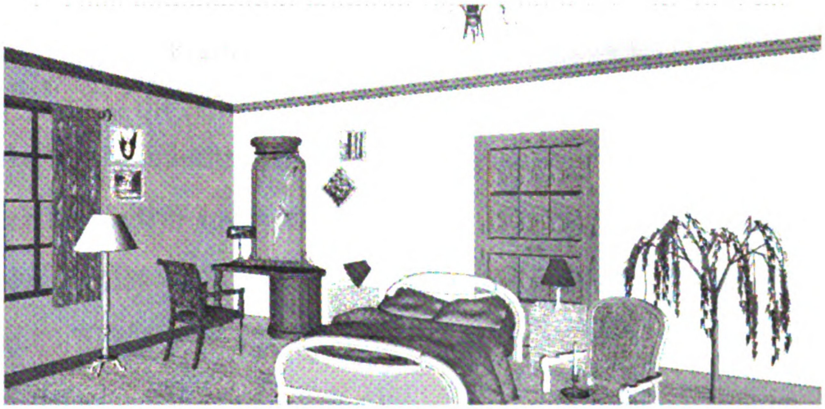


Figure 3.3. A 3D interior decoration domain

There are 13 types of entities (3D objects, e.g., chair, bed, lamp) in this domain.

3.3.2 Treasure Hunting

Figure 3.4 shows the 3D treasure hunting domain. In this domain, users walk around in a 3D castle trying to find treasures that are hidden somewhere in the rooms of a castle. Unlike the interior decoration domain where users give spoken commands to the system to move around and change decoration, in the treasure hunting domain, users walk around inside the castle and move objects by themselves, but the user has to talk to the system to get hints about where to find the treasure. Users' eye gaze fixations are recorded during the human-machine conversation.

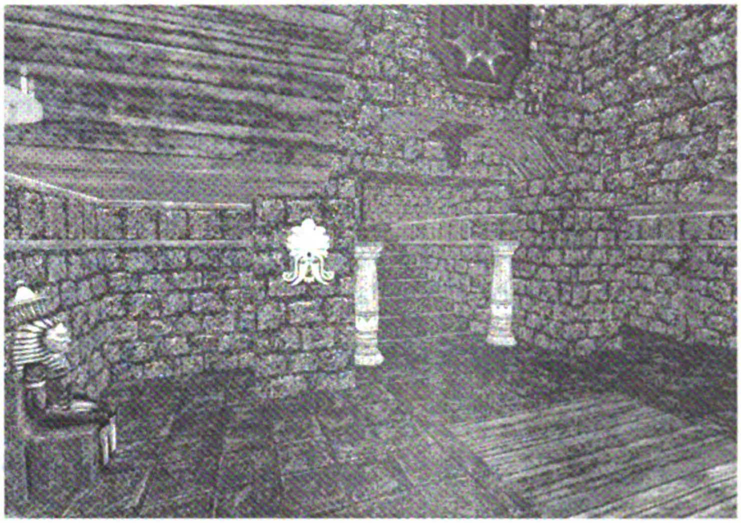


Figure 3.4. A treasure hunting domain

Compared to the interior decoration domain, the treasure hunting domain provides a richer interactive environment that involves more complex scenes and tasks, which enables studies on automatic vocabulary acquisition during human-machine conversation.

The underlying architecture supporting these two domains can be used to develop similar 3D applications such as virtual tourism guide and virtual reality personnel training.

CHAPTER 4

Incorporation of Non-verbal Modalities in Language Models for Spoken Language Processing

In multimodal conversational systems, speech recognition performance is critical in interpreting user inputs. Only after speech is correctly recognized, is the system able to further extract semantic meaning from the recognized hypothesis. Although mutual disambiguation of multiple modalities [74] can alleviate the problem with speech recognition, speech recognition is still a bottleneck to achieving robust multimodal interpretation.

This chapter presents the use of non-verbal modalities to help speech recognition in multimodal conversation. In particular, we describe a salience driven approach to incorporate the contextual information activated by deictic gesture and eye gaze in speech recognition. This approach combines gesture-based and gaze-based salience modeling with language modeling. We further describe the application of the salience driven language models in speech recognition across different stages and present evaluation results.

4.1 A Salience Driven Framework

In this section, we first introduce the notion of salience and its applications in language processing, then describe a salience driven framework for interpretation of language in multimodal conversation.

4.1.1 Salience

Salience modeling has been used in both natural language and multimodal language processing. Linguistic salience describes entities with their accessibility in a hearer's memory and their implications in language production and interpretation. Many theories on linguistic salience have been developed, including how the salience of entities affects the form of referring expression as in the Givenness Hierarchy [35] and the local coherence of discourse as in the Centering Theory [32]. Linguistic salience modeling has been used for both language generation [98] and language interpretation. Most salience-based language interpretation have focused on reference resolution [27, 42, 58].

Visual salience measures how much attention an entity attracts from a user. An entity is more salient when it attracts a user's attention more than other entities. The cause of such attention depends on many factors including user intention, familiarity, and physical characteristics of objects. For example, an object may be salient when it has some properties the others do not have, such as it is the only one that is highlighted, the only one in its size, category, or color [57]. Visual salience can also be useful in multimodal language interpretation. Studies have shown that a user's perceived salience of entities on the graphical interface can tailor the user's referring expressions and thus can be used for multimodal reference resolution [54].

4.1.2 Salience Driven Interpretation of Spoken Language in Multimodal Conversation

During multimodal conversation, a user's deictic gesture or eye gaze fixation on the graphical display indicates the user's attention and therefore indicate salient entities. The more likely is an entity selected by a gesture or eye gaze, the more salient is this entity.

We developed a salience driven framework [13] for language interpretation in multimodal conversational systems. Figure 4.1 illustrates the salience driven interpretation of speech in this framework. As shown in the figure, the user's deictic gesture or eye gaze fixation on the graphic display signals a distribution of entities that are salient at that particular time of interaction. The contextual knowledge associated with these salient objects constitutes the salient context. This salient context can be used to help speech recognition and understanding by constraining speech hypotheses.

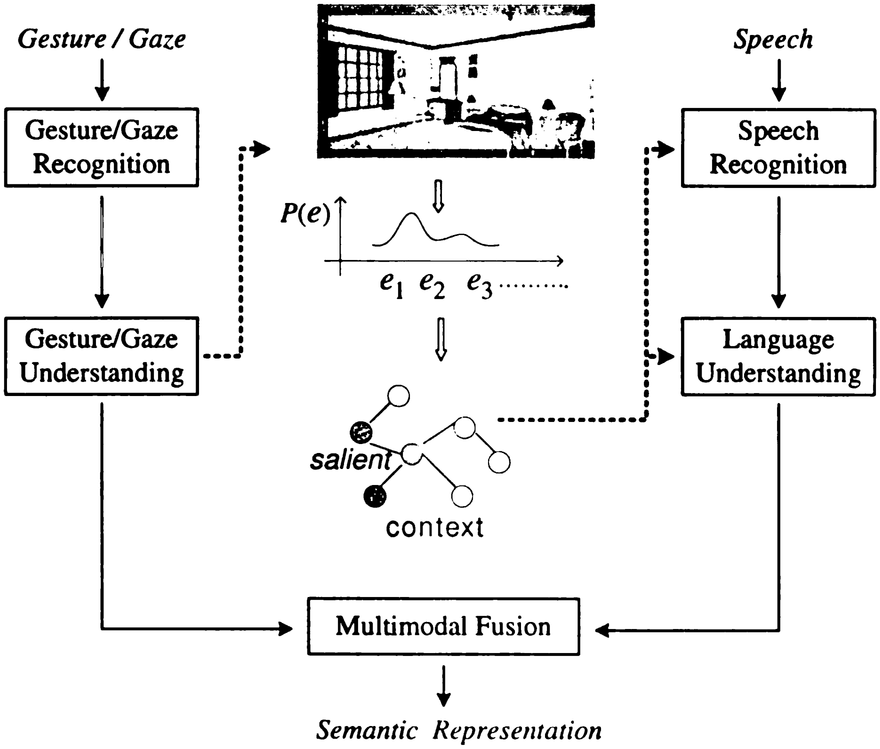


Figure 4.1. Salience driven interpretation

In this framework, there are two important operations involved: 1) the salience modeling based on gesture/gaze, and 2) the incorporation of salience information in language processing. We address these two operations in the following sections.

4.2 Gesture-Based Salience Modeling

As mentioned earlier, a deictic gesture on the graphical display can signal the underlying context that is salient at that particular time of communication. In other words, the deictic gesture will activate a salience distribution over entities in the domain. As illustrated in Figure 4.2, the salience value of an entity e at time t is calculated based on the probabilities that e is selected by the gestures $g = \{g_i\}$ occurring prior to time t.

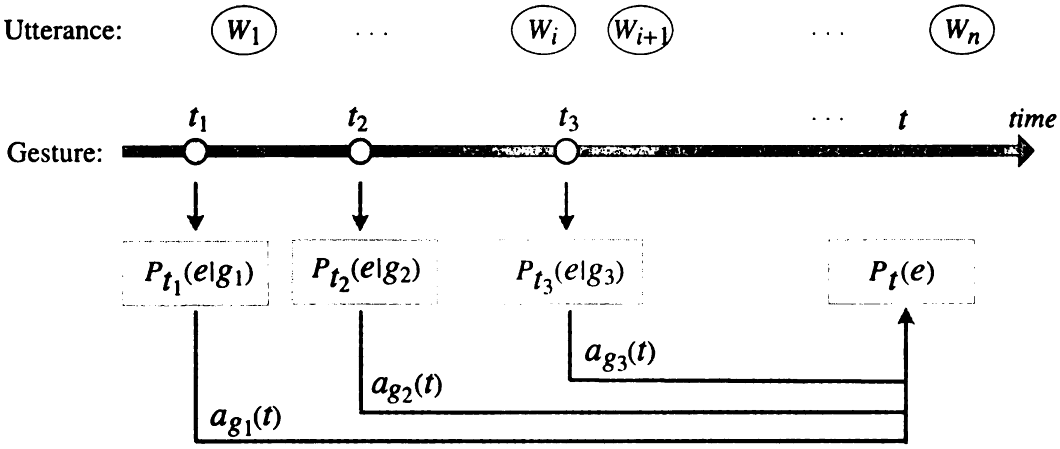


Figure 4.2. Gesture-based salience modeling

More specifically, for an entity e in the domain, its salience value at time t is calculated as follows [13]:

$$p_t(e) = \begin{cases} \dfrac{\sum\limits_{g} \alpha_g(t)p(e|g)}{\sum\limits_{e,g} \alpha_g(t)p(e|g)} & \sum\limits_{e,g} \alpha_g(t)p(e|g) \neq 0 \\ 0 & \sum\limits_{e,g} \alpha_g(t)p(e|g) = 0 \end{cases} \tag{4.1}$$

where $p(e|g)$ is the probability of entity e being selected by gesture g (calculated based on the distance from the gesture point to the center of the entity), $\alpha_g(t)$ is the weight of gesture g contributing to the salience distribution at time t.

Gesture weight $\alpha_g(t)$ is defined as follows:

$$\alpha_g(t) = \begin{cases} e^{-\frac{t-t_g}{2000}} & t \geq t_g \\ 0 & t < t_g \end{cases} \tag{4.2}$$

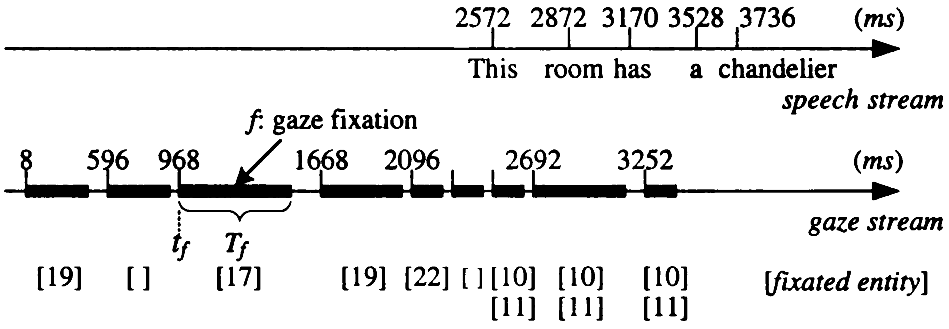
where t_g stands for the beginning time (in milliseconds) of gesture g. Weight $\alpha_g(t)$ says that gesture g has more impact on the salience distribution at a time closer to the gesture's occurrence. Note that at any time t, only gestures occurring before t (i.e., $t \geq t_g$) can contribute to the salience distribution at time t.

4.3 Gaze-Based Salience Modeling

Psycholinguistic experiments have shown that eye gaze is tightly linked to human language processing. Eye gaze is one of the reliable indicators of what a person is "thinking about" [37]. The direction of gaze carries information about the focus of the users attention [49]. The perceived visual context influences spoken word recognition and mediates syntactic processing [89, 101]. In addition, directly before speaking a word, the eyes move to the mentioned object [31].

Motivated by these psycholinguistic findings about eye gaze's link to speech, we use eye gaze information in salience models to help spoken language processing.

Figure 4.3 shows an excerpt of the speech and gaze fixation stream. In the speech stream, each word starts at a particular timestamp. In the gaze stream, each gaze fixation f has a starting timestamp t_f and a duration T_f. Gaze fixations can have different durations. An entity e on the graphical display is fixated by gaze fixation f if the area of e contains the fixation point of f. One gaze fixation can fall on multiple entities or no entity.



([10] - *bedroom*; [11] - *chandelier*; [17] - *lamp_2*; [19] - *bed_frame*; [22] - *door*)

Figure 4.3. An excerpt of speech and gaze stream data

We first define a gaze fixation set $F_{t_0}^{t_0+T}(e)$, which contains all gaze fixations that fall on entity e within a time window $t_0 \sim (t_0 + T)$:

$$F_{t_0}^{t_0+T}(e) = \{f | f \text{ falls on } e \text{ within } t_0 \sim (t_0 + T)\} \tag{4.3}$$

We model gaze-based salience in two ways [82]:

- Gaze Salience Model 1

  Salience model 1 is based on the assumption that when an entity has more gaze fixations on it than other entities, this entity is more likely attended by the user and thus has higher salience:

  $$p_{t_0,T}(e) = \frac{\#\text{elements in } F_{t_0}^{t_0+T}(e)}{\sum_e (\#\text{elements in } F_{t_0}^{t_0+T}(e))} \tag{4.4}$$

  Here, $p_{t_0,T}(e)$ tells how likely it is that the user is focusing on entity e within time period $t_0 \sim (t_0 + T)$ based on how many gaze fixations are on e among all gaze fixations that fall on entities within $t_0 \sim (t_0 + T)$.

- Gaze Salience Model 2

  Salience model 2 is based on the assumption that when an entity has longer gaze fixations on it than other entities, this entity is more likely attended by the user and thus has higher salience:

$$p_{t_0,T}(e) = \frac{D_{t_0}^{t_0+T}(e)}{\sum\limits_{e} D_{t_0}^{t_0+T}(e)} \tag{4.5}$$

where

$$D_{t_0}^{t_0+T}(e) = \sum_{f \in F_{t_0}^{t_0+T}(e)} T_f \tag{4.6}$$

Here, $p_{t_0,T}(e)$ tells how likely it is that the user is focusing on entity e within time period $t_0 \sim (t_0 + t)$ based on how long e has been fixated by gaze fixations among the overall time length of all gaze fixations that fall on entities within $t_0 \sim (t_0 + T)$.

4.4 Salience Driven Language Modeling

Given salience models, the next question is how to incorporate this salient contextual information in language processing. In this section, we describe the building of salience driven language models for speech recognition. We first give a review of the typical language models used in speech recognition, then describe how to build salience driven language models based on those baseline models.

4.4.1 Language Models for Speech Recognition

The task of speech recognition is to, given an observed spoken utterance O, find the word sequence W^* such that

$$W^* = \arg\max_{W} p(O|W)p(W) \tag{4.7}$$

where $p(O|W)$ is the acoustic model and $p(W)$ is the language model.

In speech recognition systems, the acoustic model provides the probability of observing the acoustic features given hypothesized word sequences, and the language model provides the prior probability of a sequence of words. The language model is represented as follows:

$$p(W) = p(w_1^n) = p(w_1)p(w_2|w_1)p(w_3|w_1^2)\ldots p(w_n|w_1^{n-1}) \tag{4.8}$$

The language model can be approximated by a bigram model using first-order Markov assumption:

$$p(w_1^n) = \prod_{k=1}^{n} p(w_k|w_{k-1}) \tag{4.9}$$

or by a trigram model using second-order Markov assumption:

$$p(w_1^n) = \prod_{k=1}^{n} p(w_k|w_{k-1}, w_{k-2}) \tag{4.10}$$

By clustering words into classes, the class-based n-gram model reduces the training data requirement and improves the robustness of probability estimates compared to the word n-gram model. The class-based bigram model is given by [6]:

$$p(w_i|w_{i-1}) = p(w_i|c_i)p(c_i|c_{i-1}) \tag{4.11}$$

where c_i and c_{i-1} are the classes of word w_i and w_{i-1} respectively.

Probabilistic context free grammar (PCFG) can also be used as a language model in speech recognition by constraining the speech recognizer to generate only grammatical sentences as defined by the grammar.

4.4.2 Salience Driven N-Gram Models

Statistical n-gram models are widely used in speech recognition. We incorporate the gesture/gaze-based salience modeling into the bigram model and the class-based bigram model to build salience driven n-gram models [13,81] for speech recognition.

- Salience driven bigram model

  The salience driven bigram probability $p_s(w_i|w_{i-1})$ is given by:

$$p_s(w_i|w_{i-1}) = \frac{p(w_i|w_{i-1}) + \lambda \sum_e p(w_i|w_{i-1}, e)p_t(e)}{1+\lambda} \tag{4.12}$$

where $p_t(e)$ is the salience distribution, as modeled in equation (4.1), λ is the priming weight. The priming weight λ decides how much the original bigram probability will be tailored by the salient entities that are indicated by gestures. Currently, we set $\lambda = 2$ empirically. We also tried to learn the priming weight by an EM algorithm. However, we found out that the learned λ performed worse than the empirical one in our experiments. This is partially due to insufficient development data. Bigram probabilities $p(w_i|w_{i-1})$ were estimated by the maximum likelihood estimation using Katz's backoff method [51] with frequency cutoff of 1. The same method was used to estimate $p(w_i|w_{i-1}, e)$ from the users' speech transcripts with entity annotation of e.

- Salience driven class-based bigram model

  The salience driven class-based bigram probability $p_s(w_i|w_{i-1})$ is given by:

$$p_s(w_i|w_{i-1}) = \begin{cases} p(c_i|c_{i-1}) \sum_e p(w_i|c_i, e)p_t(e) & \sum_e p_t(e) \neq 0 \\ p(w_i|w_{i-1}) & \sum_e p_t(e) = 0 \end{cases} \tag{4.13}$$

  where $p_t(e)$ is the salience distribution, c_i and c_{i-1} are the semantic classes of words w_i and w_{i-1} respectively, $p_s(w_i|c_i, e)$ is learned with maximum likelihood estimation from the utterances talking about entity e.

4.4.3 Salience Driven PCFG

Building salience driven PCFG [81] as language model includes three steps: 1) construct a context free grammar (CFG) specific to the application domain; 2) for each entity in the domain, train entity-specific PCFG based on the utterances talking about that particular entity; 3) create salience driven PCFG based on the entity salience distribution and entity-specific PCFGs.

More specifically, we build salience driven PCFG for the 3D interior decoration domain (Section 3.3.1) as follows. Based on the domain knowledge, we first define a domain-specific CFG as shown in Figure 4.4. This CFG covers all the language that is "legal" in the interior decoration domain. An utterance is said to be"legal" in the domain if a semantic representation specific to the domain can be built from the utterance. The defined grammar covers the "legal" commands like "*this table*", "*remove this chair*", "*move this plant on this table*", and query questions like "*how much is this table?*", "*who is the artist of this painting?*", "*what is the wattage of this lamp?*".

<table>
<tr><td>S</td><td>→</td><td>NP | VP | WRB JJ VBZ NP | WRB JJ NN VBZ NP VB | WP VBZ NP PP | WRB VBZ NP VBN | VBZ NP NP</td></tr>
<tr><td>VP</td><td>→</td><td>VB NP | VB NP PP | VB NP JJ | VB NP RB</td></tr>
<tr><td>NP</td><td>→</td><td>NN | DT NN | PRP</td></tr>
<tr><td>PP</td><td>→</td><td>IN DT NN | TO DT NN</td></tr>
<tr><td>WP</td><td>→</td><td>what | who</td></tr>
<tr><td>WRB</td><td>→</td><td>how | where</td></tr>
<tr><td>JJ</td><td>→</td><td>big | black | blue | dark | expensive | gray | green | ...</td></tr>
<tr><td>VBZ</td><td>→</td><td>does | is</td></tr>
<tr><td>VB</td><td>→</td><td>add | align | bring | buy | change | delete | ...</td></tr>
<tr><td>RB</td><td>→</td><td>back | backward | backwards | down | forward | here | ...</td></tr>
<tr><td>NN</td><td>→</td><td>age | alternative | artist | artwork | back | bar | bed ...</td></tr>
<tr><td>DT</td><td>→</td><td>a | an | that | the | these | this | those</td></tr>
<tr><td>PRP</td><td>→</td><td>it | them</td></tr>
<tr><td>IN</td><td>→</td><td>about | above | against | among | around | at | behind ...</td></tr>
<tr><td>TO</td><td>→</td><td>to</td></tr>
<tr><td>VBN</td><td>→</td><td>made | produced</td></tr>
</table>

Figure 4.4. Context free grammar for the 3D interior decoration domain

We build the entity-specific PCFGs by first using the Stanford Parser [55] to parse users' transcribed utterances, then for each entity *e* in the domain, training a PCFG with maximum likelihood estimation based on the utterances talking about entity *e*. In the trained PCFG, only the lexicon-part rules are associated with probabilities. An example of trained PCFG for entity *lamp* is shown in Figure 4.5. The PCFG in Figure 4.5 is in the Java Speech Grammar Format (JSGF) and the numbers in the "/ /" are the weights of the rules. When normalized, the weights are the rule probabilities. As we can see in Figure 4.5, the words closely related to entity *lamp* such as "lamp" and "wattage" achieve higher weights in the trained PCFG. It means that those words closely related to *lamp* will be more likely chosen during the speech recognition process when the entity *lamp* is salient.

```
<S>     =  <NP> | <VP> | <WRB> <JJ> <VBZ> <NP> | ...;
<VP>    =  <VB> <NP> | <VB> <NP> <PP> | <VB> <NP> <JJ> |
           <VB> <NP> <RB>;
<NP>    =  <NN> | <DT> <NN> | <PRP>;
<PP>    =  <IN> <DT> <NN> | <TO> <DT> <NN>;
<DT>    =  /117/ this | /59/ the | /16/ that | /3/ these | /1/ those |
           /1/ a | /1/ an;
<IN>    =  /34/ of | /17/ on | /10/ about | /7/ with | /4/ in |
           /2/ behind | ...;
<JJ>    =  /8/ many | /2/ much | /1/ small | /1/ left | /1/ expensive | ...;
<NN>    =  /144/ lamp | /24/ wattage | /7/ place | /7/ information |
           /6/ table | ...;
<PRP>   =  /3/ it | /1/ them;
<RB>    =  /9/ here | /2/ back | /2/ up | /2/ there;
<TO>    =  to;
<VB>    =  /27/ remove | /18/ move | /7/ show | /6/ put |
           /6/ change | ...;
<VBN>   =  /2/ made | /1/ produced;
<VBZ>   =  /30/ is | /3/ does;
<WP>    =  /26/ what | /4/ who;
<WRB>   =  /9/ how | /5/ where;
```

Figure 4.5. Trained PCFG for entity *lamp* in the 3D interior decoration domain

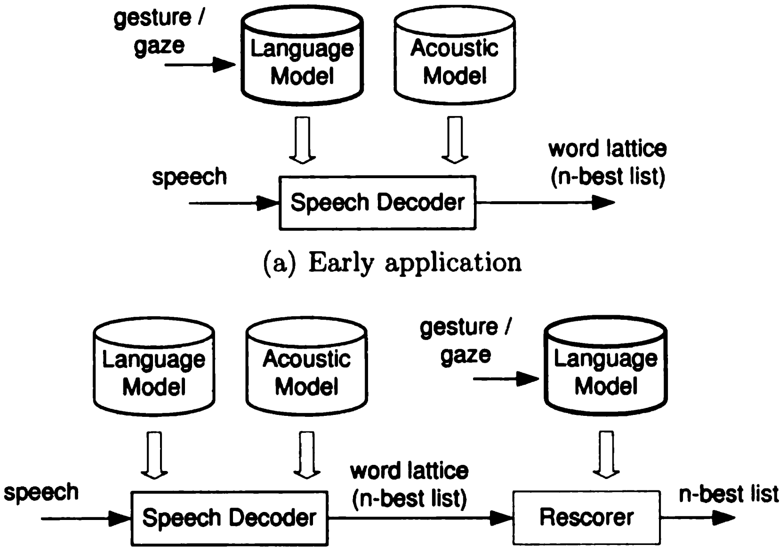
Given entity-specific PCFGs, salience driven PCFG is created by combining the PCFGs associated with the salient entities. The weight of a rule r in the salience driven PCFG is given by:

$$w(r) = \sum_{e} w_e(r)p(e) \tag{4.14}$$

where $p(e)$ is the salience distribution, $w_e(r)$ is the weight of rule r in the PCFG specific to entity e.

4.5 Application of Salience Driven Language Models for ASR

The salience driven language models can be integrated into speech recognition in two stages: an early stage before word lattice (n-best list) is generated, or in a later stage where the word lattice (n-best list) is post-processed (Figure 4.6).



(a) Early application

(b) Late application

Figure 4.6. Application of salience driven language model in speech recognition

4.5.1 Early Application

For the early application, as Figure 4.6(a) shows, salience driven language model is used together with the acoustic model to generate the word lattice, typically by Viterbi search.

Compared to n-gram models, CFG-based language models put more strict constraint on the speech recognition process, specifically on choosing the next set of possible words following a path during the searching process. When an n-gram model is used, the next set of possible words includes any words in the vocabulary with non-zero transition probabilities (as specified by the n-gram model) from the previous n-1 words along the path. When a CFG-based language model is used, the next set of possible words only includes those allowable words as defined by the grammar.

4.5.2 Late Application

For the late application, as shown in Figure 4.6(b), the salience driven n-gram language model is used to rescore the word lattice generated by a speech recognizer with a basic language model not involving salience modeling. A word lattice consists of a list of nodes and edges (Figure 4.7). In the word lattices, each node represents a word hypothesis and each edge represents a word transition. Each path going from the start node <s> to the end node </s> forms a sentence recognition hypothesis. Given a word lattice, A* search can be applied to find the n-best paths in the word lattice.

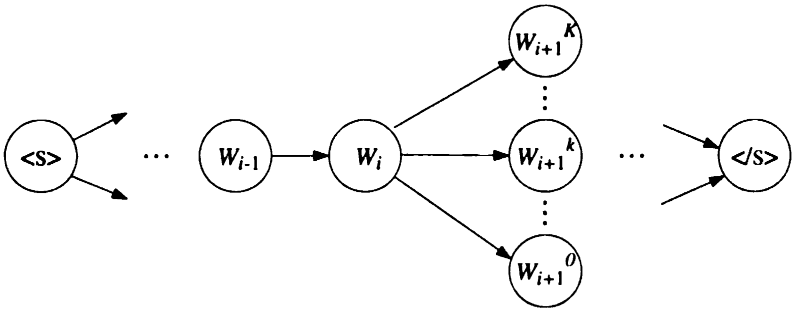


Figure 4.7. A* search in word lattice

A* search finds in a graph the optimal path from a given initial node to a given goal node. Specifically, in the word lattice shown in Figure 4.7, the task of A* search is to find a path from sentence start node "<s>" to sentence end node "</s>" that has the highest score. The score of a path $L = (w_0, w_1, \ldots, w_n)$ is defined as

$$f(L) = \sum_{i=0}^{n} (\log p_a(w_i) + \log p(w_i|w_{i-1})) \tag{4.15}$$

where $p_a(w_i)$ is the acoustic model probability and $p(w_i|w_{i-1})$ is the language model probability. The language model probabilities can be tailored by the salience driven language models described in Section 4.4.2.

In the word lattice, each node (i.e., a word hypothesis) is associated with a score. The score of a word w_i depends on two parts: the *true score* $g(w_i)$ that measures the actual score of the path from the start node to the current node, and the *heuristic score* $h(w_i)$ that measures the expected score of the path from the current node to the goal node. In each step of the A* search, the next node to expand is chosen as the one with the highest score $(g(w_i) + h(w_i))$ among the ending nodes of all previous partial paths that have been explored.

Before A* search begins, the heuristics at each node w_i are first calculated:

$$h(w_i) = \max_k \{h(w_{i+1}^k) + \log p_a(w_{i+1}^k) + \log p(w_{i+1}^k | w_i)\} \tag{4.16}$$

where $h(</s>) = 0$.

During the A* searching process, the score of the path up to node w_i is calculated as:

$$g(w_i) = g(w_{i-1}) + \log p_a(w_i) + \log p(w_i | w_{i-1}) \tag{4.17}$$

where $g(<s>) = 0$.

A late application of gaze-tailored language model was reported in [21], where the language model tailored by eye gaze was used to directly reorder the n-best list of speech recognition to get better 1-best recognition. We will show in Section 4.6.5 that the early application works better than the late application.

4.6 Evaluation

In the 3D interior decoration domain, we empirically evaluate the different salience driven language models when applied at the two stages for speech recognition.

4.6.1 Speech and Gesture Data Collection

We conducted a wizard-of-oz study to collect speech and gesture data for our evaluation using the system described in Chapter 3. In the study, users were asked to accomplish two tasks. Task 1 was to clean up and redecorate a messy room. Task 2 was to arrange and decorate the room so that it looks like the room in the pictures provided to the user. Each of these tasks put the user into a specific role (e.g., college student, professor, etc.), and the task had to be completed with a set of constraints (e.g. budget of furnishings, bed size, number of domestic products, etc.). A detailed description of the user study in the interior decoration domain is given in Appendix A.1.

From 5 users' interactions with the system, we collected 649 utterances with accompanying gestures. The vocabulary size of the collected utterances is 250 words.

Each utterance was transcribed and annotated with referred entities. For example, an utterance like "*remove this lamp*" accompanied by a deictic gesture was annotated with the true entity *lamp1* as indicated by the gesture, while an utterance like "*move this lamp to this table*" accompanied by two deictic gestures were annotated with the entities *lamp1* and *table1* as indicated by the two gestures respectively.

Each gesture results in a set of possibly selected entities. The selection probabilities of the entities are calculated based on the distances from the gesture point to the center of the entities.

All the collected data, together with the speech transcripts and entity annotation, are saved in XML format. Figure 4.8 shows an excerpt form one of the XML data files. The excerpt is the record of one turn in the conversation between the system and one user. In this turn, the user pointed to the entity *picture_girl* and said "*flip this picture one hundred eighty degrees*". The pointing gesture resulted in an ambiguous selection of three entities (*bedroom*, *picture_girl*, *table_pc*) with different probabilities.

4.6.2 Evaluation Results on Speech and Gesture Data

We compare the performances of the following different language models trained in our domain:

```
<turn>
 <user_input>
  <gesture>
   <curve start="2153" end="2309">
    <point>613 183</point>
    <point>613 183</point>
   </curve>
   <selection>
    <entity text="bedroom">0.458000</entity>
    <entity text="picture_girl">0.530700</entity>
    <entity text="table_pc">0.011300</entity>
   </selection>
  </gesture>
  <speech>
   <entity_annotation>
    picture_girl
   </entity_annotation>
   <transcription>
    flip this picture one hundred eighty degrees
   </transcription>
   <waveform>2005916-144311-707.wav</waveform>
  </speech>
 </user_input>
</turn>
```

Figure 4.8. An Excerpt of XML Data File

- Standard bigram model (Bigram)

- Standard trigram model (Trigram)

- Standard class-based bigram model (C-Bigram)

- Salience driven bigram model (S-Bigram)

- Salience driven class-based bigram model (S-C-Bigram)

- Standard PCFG (PCFG)

- Salience driven PCFG (S-PCFG)

The evaluation metrics include the following aspects related to recognition results:

- Word error rate of the best hypothesis (WER)

- Word lattice WER (Lattice-WER)

  The minimal WER of all possible paths through the word lattice (output of speech recognition).

Since we are building a conversational system, we are also interested in the following metrics related to semantic interpretation:

- Concept identification precision (CI-Precision)

  The percentage of correctly identified concepts out of the total number of concepts in the 1-best recognition hypothesis.

- Concept identification recall (CI-Recall)

  The percentage of correctly identified concepts out of the total number of concepts in a user's utterance (speech transcript).

- F-measurement (F-score)

  $$F = \frac{(\beta^2 + 1) \times \text{CI-Precision} \times \text{CI-Recall}}{\beta^2 \times \text{CI-Precision} + \text{CI-Recall}}$$

  where $\beta = 1$ in this experiment.

The evaluation was done by an eight-fold cross validation. We compare the performances of the salience driven language models for both early and late applications.

Results of Early Application

Table 4.1 shows the experimental results of the early application of different language models on the utterances with accompanying gestures. Among the n-gram models, the performance of the trigram model is roughly the same as the bigram model. The salience driven bigram (S-Bigram) model improved speech recognition and understanding compared to the three baselines (Bigram, Trigram, and C-Bigram). Compared to the best baseline of the trigram model, the S-Bigram model reduced the WER by 7%. A t-test showed that this was a significant change: $t = 3.38$, $p < 0.004$.

Language Model	Lattice-WER	WER	CI-Precision	CI-Recall	F-score
Bigram	0.250	0.321	0.830	0.793	0.811
Trigram	0.258	0.312	0.838	0.797	0.817
C-Bigram	0.292	0.371	0.856	0.748	0.798
S-Bigram	**0.243**	**0.291**	0.861	0.830	0.845
S-C-Bigram	0.412	0.448	**0.863**	0.623	0.724
PCFG	0.323	0.360	0.819	0.816	0.817
S-PCFG	0.319	0.355	0.862	**0.845**	**0.853**

Table 4.1. Performances of the early application of different language models on speech-gesture data

The S-Bigram model increased the precision and recall of concept identification by 3% and 4% respectively. The overall F-measurement achieved by the S-Bigram model gained an increase of 3%. A t-test showed that this was also a significant improvement: $t = 3.01$, $p < 0.002$. The S-C-Bigram model achieved the best result on the precision of concept identification, but had the worst results on all other metrics.

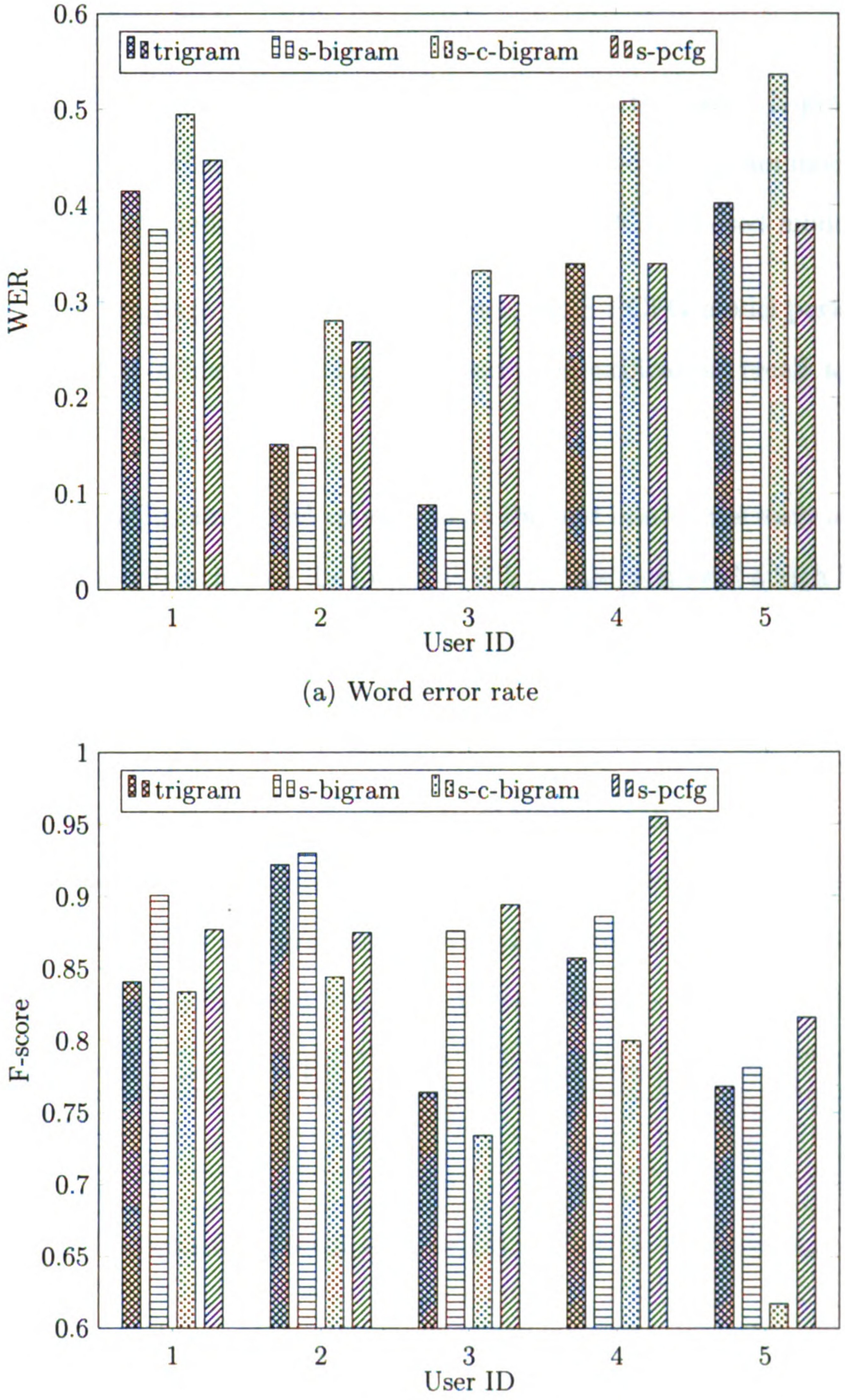
Comparing class-based n-gram models (C-Bigram, S-C-Bigram) to n-gram models (Bigram, Trigram, S-Bigram), we can see that class-based n-gram models achieve better concept identification precision but worse concept identification recall and WER. The performances of the class-based n-gram models depend on how the classes of words are defined. When one unique class is defined for each unique word, there will be no difference between n-gram models and class-based n-gram models. In our experiment, we define different classes for the words with key semantic concepts, whereas a single class is assigned to all other words. With this class definition, the class-based bigram models contain n-gram probability information about the words with key semantic concepts but lost the information for the non-key words with one same class. Therefore, using the class-based n-gram models in speech recognition, it is hard to correctly recognize the non-key words with one same class, whereas the words with key semantic concepts are more likely to appear in the recognition result, though many of them incorrectly recognized. This leads to a better concept identification precision but worse concept identification recall and WER.

Compared to the standard PCFG model, the salience driven PCFG (S-PCFG) model increased the precision and recall of concept identification by 5% and 3.5% respectively. The overall F-measurement was increased by 4%. A t-test confirmed that this was a significant improvement: $t = 3.30$, $p < 0.001$. The S-PCFG model did not change the WER much compared to the standard PCFG model. A t-test confirmed that this change in WER was not significant.

When compared to the trigram model, the S-PCFG model did not improve the WER but improved the language understanding. The F-measurement was increased by 4%. A t-test showed that this was a significant improvement: $t = 2.77$, $p < 0.003$. The worse WER of the S-PCFG model is due to the lesser flexibility of grammar-based language models than n-gram language models. Grammar-based language models place too much constraint on what language can be recognized, which hurts the recognition of complex utterances. On the other hand, after salience tailoring, the more strict constraints on what words of key semantic concepts can be recognized for the salient entity makes the S-PCFG model achieve better language understanding performance than the n-gram model.

We also show the experimental results for individual users. Figure 4.9 compares the performances of different salience driven language models in early application for each user. From the results for individual users, we can see that for most users, the performances of different salience driven language models are consistent. Compared to the best baseline of the trigram model, the S-Bigram model achieved lower WER and higher F-score for each user. The S-C-Bigram model did not show improvement over the trigram model for all users. The S-PCFG model showed its merit on improving language understanding by achieving higher F-scores than the baseline for all users except user 2. And for 3 of the 5 users, the S-PCFG model achieved the best language understanding among all different language models.

Overall, the results of the early application of the gesture-based salience driven



(a) Word error rate

(b) F-score

Figure 4.9. Performance of the early application of LMs on speech-gesture data of individual users

language models show that:

- In terms of WER, the S-Bigram model performed the best. N-gram models performed better than class-based n-gram models, and all n-gram models except the S-C-Bigram model performed better than the PCFG-based models.

- In terms of language understanding metrics, the S-PCFG model performed the best in that it achieved the highest concept identification recall and overall F-measurement.

- Overall, the S-Bigram model appears to be the best one for the early application in that it not only achieved the lowest WER but also achieved a high F-score on concept identification (close to the highest one achieved by the S-PCFG model).

Results of Late Application

We further compared different n-gram models: C-Bigram, S-Bigram, and S-C-Bigram during the late application. In these experiments, the standard trigram model trained on our domain was first used to generate word lattices, then the salience driven models were used in A* search (Section 4.5.2) to find the best paths in the word lattices.

Language Model	Lattice-WER	WER	CI-Precision	CI-Recall	F-score
C-Bigram	0.258	0.334	0.831	0.784	0.807
S-Bigram	0.258	**0.294**	0.854	**0.834**	**0.844**
S-C-Bigram	0.258	0.316	**0.858**	0.786	0.821

Table 4.2. Performance of the late application of LMs on speech-gesture data

Table 4.2 shows the results of the three models on the utterances with accompanying gestures. In the late application, the S-Bigram model performed the best with the exception of concept identification precision. Compared to the trigram model, the S-Bigram in late application decreased the WER by 6%. A t-test showed that this was a significant change: $t = 2.66$, $p < 0.005$. On language understanding, the

S-Bigram model increased the F-measurement by 3% compared to the trigram model. A t-test confirmed that this was a significant improvement: $t = 2.92$, $p < 0.002$.

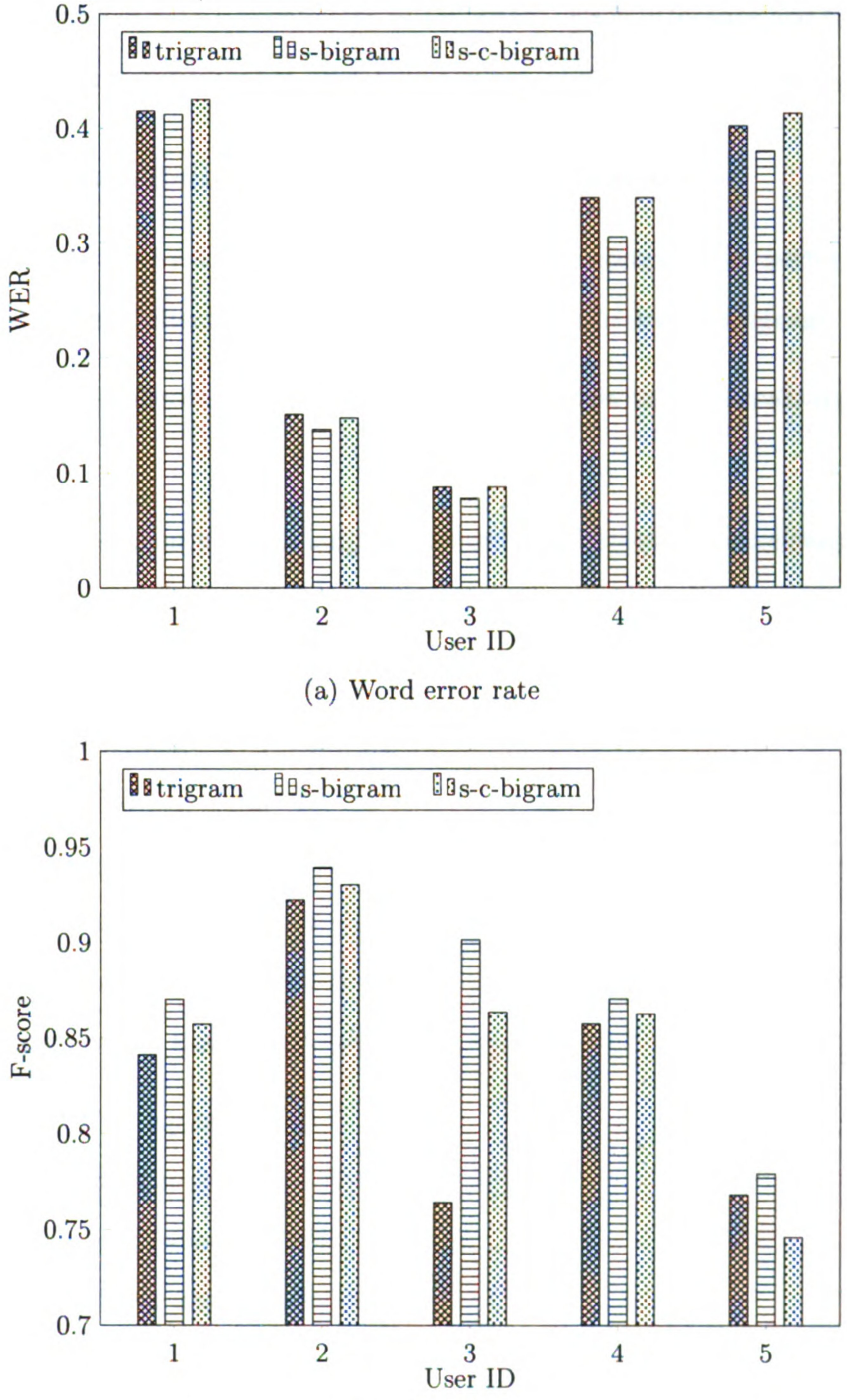
Compared to Table 4.1, Table 4.2 shows that there is no difference in performance whether the S-Bigram model is applied early or later. However, a significant difference is observed for the S-C-Bigram model. The S-C-Bigram model performed much better when it was applied in a later stage. However, its performance was close to the baseline (trigram model). The WER change achieved by the S-C-Bigram model was not statistically significant from the t-test ($t = 0.94$, N.S.), neither was the F-measurement ($t = 0.22$, N.S.).

The experimental results of the late application of the three n-gram models for individual users are shown in Figure 4.10. The results demonstrate the consistency of the performances of different salience driven language models in late application for most users. Compared to the baseline of the trigram model, the S-Bigram model improved both speech recognition and language understanding when applied in a late stage. The S-C-Bigram model did not improve speech recognition either when applied in a late stage, but it improved language understanding for most of the users. Compared to its performance on speech recognition in early application, the S-C-Bigram model performed better in late application for all the users.

4.6.3 Speech and Eye Gaze Data Collection

We conducted user studies to collect speech and eye gaze data. In the experiments, a static 3D bedroom scene was shown to the user. The system verbally asked the user a list of questions one at a time about the bedroom and the user answered the questions by speaking to the system. A detailed description of the user study is given in Appendix A.2.

The user's speech was recorded through an open microphone and the user's eye gaze was captured by an Eye Link II eye tracker. From 7 users' experiments, we col-



(a) Word error rate

(b) F-score

Figure 4.10. Performance of the late application of LMs on speech-gesture data of individual users

lected 554 utterances with a vocabulary of 489 words. Each utterance was transcribed and annotated with entities that were being talked about in the utterance.

4.6.4 Evaluation Results on Speech and Eye Gaze Data

Evaluation was done by a 14-fold cross validation. We compare the performances of the early and late applications of two gaze-based salience driven language models:

- S-Bigram1 – salience driven language model based on salience modeling 1 (Equation (4.4))
- S-Bigram2 – salience driven language model based on salience modeling 2 (Equation (4.5))

Table 4.3 and Table 4.4 show the results of the early and late applications of the salience driven language models based on eye gaze. We can see that all word error rates (WERs) are high. In the experiments, users were instructed to only answer systems questions one by one. There was no flow of a real human-machine conversation. In this setting, users were more free to express themselves than in the situation where users believed they were conversing with a machine. Thus, we observe much longer sentences that often contain disfluencies. Here is one example:

> System: "*How big is the bed?*"
> User: "*I would to have to offer a guess that the bed, if I look the chair that's beside it [pause] in a relative angle to the bed, it's probably six feet long, possibly, or shorter, slightly shorter.*"

The high WER was mainly caused by the complexity and disfluencies of users' speech. Poor speech recording quality is another reason for the bad recognition performance. It is found that the trigram model performed worse than the bigram model in the experiment. This is probably due to the sparseness of trigrams in the corpus. The amount of data available is too small considering the vocabulary size.

Language Model	Lattice-WER	WER
Bigram	0.613	0.707
Trigram	0.643	0.719
S-Bigram 1	0.605	0.690
S-Bigram 2	0.604	0.689

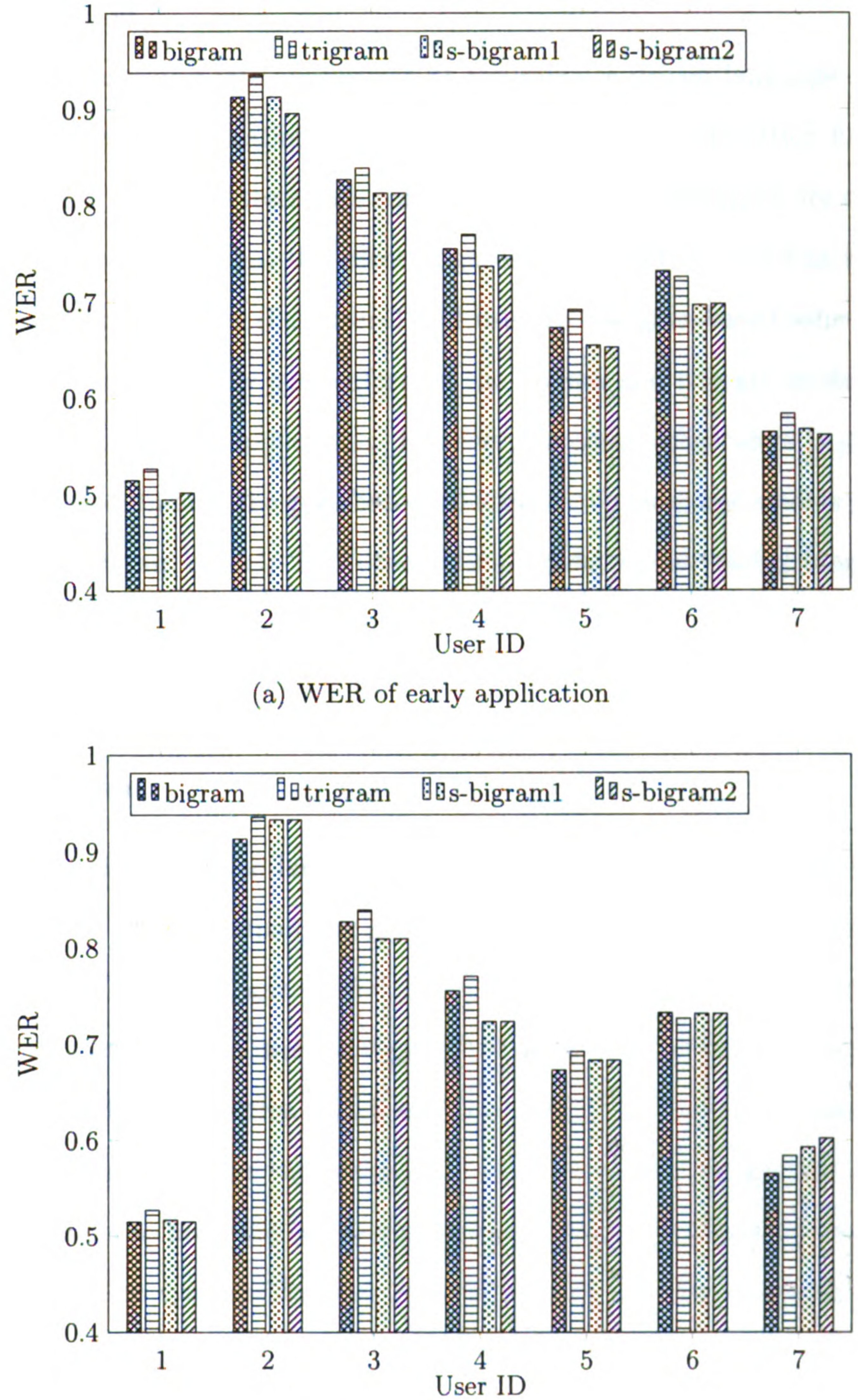
Table 4.3. WER of the early application of LMs on speech-gaze data

Language Model	Lattice-WER	WER
S-Bigram 1	0.643	0.709
S-Bigram 2	0.643	0.710

Table 4.4. WER of the late application of LMs on speech-gaze data

The S-Bigraml and S-Bigram2 achieved similar results in both early application (Table 4.3) and late application (Table 4.4). In early application, the S-Bigraml model performed better than the trigram model (t = 5.24, $p < 0.001$) and the bigram model (t = 3.31, $p < 0.001$). The S-Bigram2 model also performed better than the trigram model (t = 5.15, $p < 0.001$) and the bigram model (t = 3.33, $p < 0.001$) in early application. In late application, the S-Bigraml model performed better than the trigram model (t = 2.11, $p < 0.02$), so did the S-Bigram2 model (t = 1.99, $p < 0.025$). However, compared to the bigram model, the S-Bigraml model did not change the recognition performance significantly in late application, neither did the S-Bigram2 model.

We also compare performances of the salience driven language models for individual users. In early application (Figure 4.11a), both the S-Bigraml and the S-Bigram2 model performed better than the baselines of the bigram and trigram models for all users except user 2 and user 7. T-tests have shown that these are significant improvements. For user 2, the S-Bigraml model achieved the same WER as the bigram model. For user 7, neither of the salience driven language models improved recognition compared to the bigram model. In late application (Figure 4.11b), only for user 3 and user 4, both salience driven language models performed better than the baselines of the bigram and trigram models. These improvements have also been confirmed by



(a) WER of early application

(b) WER of late application

Figure 4.11. WERs of application of LMs on speech-gaze data of individual users

t-tests as significant.

Comparing early and late application of the salience driven language models, it is observed that early application performed better than late application for all users except user 3 and user 4. T-tests have confirmed that these differences are significant.

It is interesting to see that the effect of gaze-based salience modeling is different among users. For two users (i.e., user 3 and user 4), the gaze-based salience driven language models consistently out-performed the bigram and trigram models in both early application and late application. However, for some other users (e.g., user 7), this is not the case. In fact, the gaze-based salience driven language models performed worse than the bigram model. This observation indicates that during language production, a user's eye gaze is voluntary and unconscious. This is different from deictic gesture, which is more intentionally delivered by a user.

4.6.5 Discussion

Gesture-based salience driven language models are built on the assumption that the entity selected by the accompanying gesture of a user's utterance is the topic of the user's utterance. Similarly, gaze-based salience driven language models are built on the assumption that when a user's eye gaze is fixating on an entity, the user is saying something related to the entity. With this assumption, gesture/gaze-based salience driven language models have the potential to improve speech recognition by biasing the speech decoder to favor the words that are consistent with the entity indicated by the user's gesture or eye gaze fixation, especially when the user's utterance contains words describing unique characteristics of the object. These particular characteristics could be the object's name or physical properties (e.g., color, material, size).

An example where the gesture-based salience driven language model helped speech recognition is shown in Figure 4.12. In this example, a user pointed to the entity *table_square* in the bedroom scene and said "*show me details on this desk*". The

Utterance: "show me details on this desk"

Gesture selection:
p(bedroom) = 0.0050
p(lamp_floor) = 0.1954
p(couch_mrsofa) = 0.1409
p(lamp_floor2) = 0.0510
p(table_square) = 0.6077

Bigram n-best list:	**S-Bigram n-best list**:
show me details on this bed	*show me details on this desk*
show me details on this desk	*show me details on this bed*
show me details on this back	*show me details on this back*
show me details on that's desks	*show me details on this desk a*
show me details on that's desk	*show me details on that's desk*
show me details on that's that's	*show me details on that's desk a*

Figure 4.12. N-best lists of speech recognition for utterance "*show me details on this desk*"

user's gesture resulted in a set of candidate entities being selected, in which the correct one (i.e., *table_square*) was assigned the highest selection probability of 0.6077. Two n-best lists, the bigram n-best list and S-Bigram n-best list, were generated by the speech recognizer when the standard bigram model and the salience driven bigram model were applied respectively. When the standard bigram model was applied, the speech recognizer did not get the correct recognition. When the salience driven bigram model was applied, the speech recognizer recognized the user's utterance correctly.

Figures 4.13 and 4.14 show the word lattices of the utterance generated by the speech recognizer using the standard bigram model and the salience driven bigram model respectively. The n-best lists in Figure 4.12 were generated from those word lattices. In the word lattices, each path going from the start node <s> to the end node </s> forms a recognition hypothesis. The bigram probabilities along the edges are in the logarithm of base 10. In the standard bigram case, although the probability of bigram "this desk" (-1.3952) is slightly higher than the probability of "this bed" (-1.4380), the speech recognizer got the wrong recognition, i.e., the correct speech

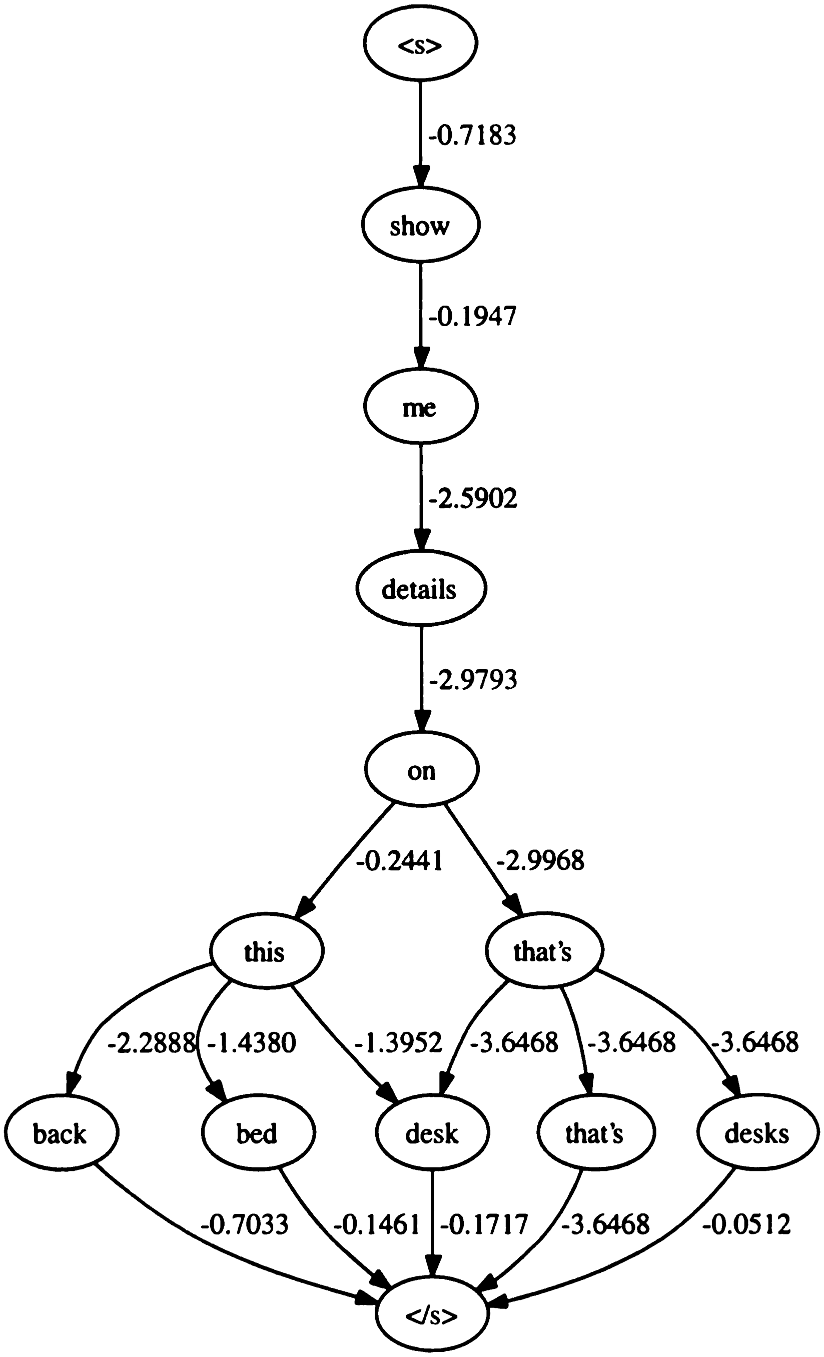


Figure 4.13. Word lattice of utterance "*show me details on this desk*" generated by using **standard bigram model**

recognition hypothesis is not the first one in the n-best list (Figure 4.12). This is because the system tries to find an overall best speech recognition hypothesis by considering both language confidence and acoustic confidence. After tailoring the

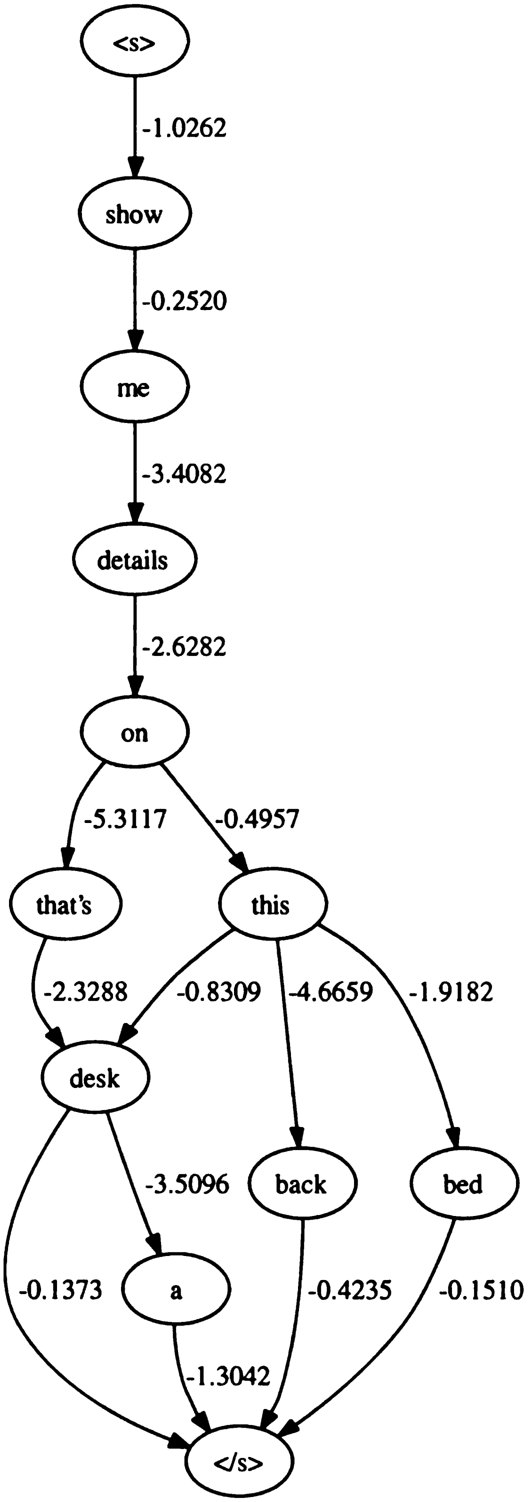


Figure 4.14. Word lattice of utterance "*show me details on this desk*" generated by using **salience driven bigram model**

standard bigram model with gesture selection, in the resulting salience driven bigram model, the probability of bigram "this desk" is increased (-0.8309) while the probability of "this bed" is decreased (-1.9182). This enlarged bigram probability difference ensures that "this desk" is on the overall best speech hypothesis generated by the speech recognizer with the salience driven language model.

Utterance: "move the red chair over here"

Gesture selection:
$p(\text{bedroom}) = 0.0001$
$p(\text{curtains\_1}) = 0.0061$
$p(\text{table\_pc}) = 0.2229$
$p(\text{chair\_1}) = 0.7196$
$p(\text{lamp\_floor}) = 0.0512$

Bigram n-best list:	**S-Bigram n-best list**:
move the rid chair over here	*move the red chair over here*
move the rid chair over here a	*move the red chair over here a*
move the rid chair over here i	*move the red chair over here i*
move the rid chair over here the	*move the red chair over here the*
move the rid chair over here it	*move the red chair over here it*
⋮	⋮

Figure 4.15. N-best lists of speech recognition for utterance "*move the red chair over here*"

Figure 4.15 shows another example where the salience driven language model helped recognize an utterance that referred visual properties of an entity. In this example, the user pointed to a red chair and then pointed to a location while saying "*move the red chair over here*". In the resulting gesture selections, the truly selected entity *chair_1* was assigned the highest probability. As shown in the bigram n-best list and the S-Bigram n-best list, the speech recognizer with the standard bigram model did not get the correct recognition result while the one with the salience driven bigram model recognized the user's utterance correctly.

The word lattices of the utterance are shown in Figures 4.16 and 4.17. In the standard bigram case, as shown in Figure 4.16, the probability of bigram "rid chair"

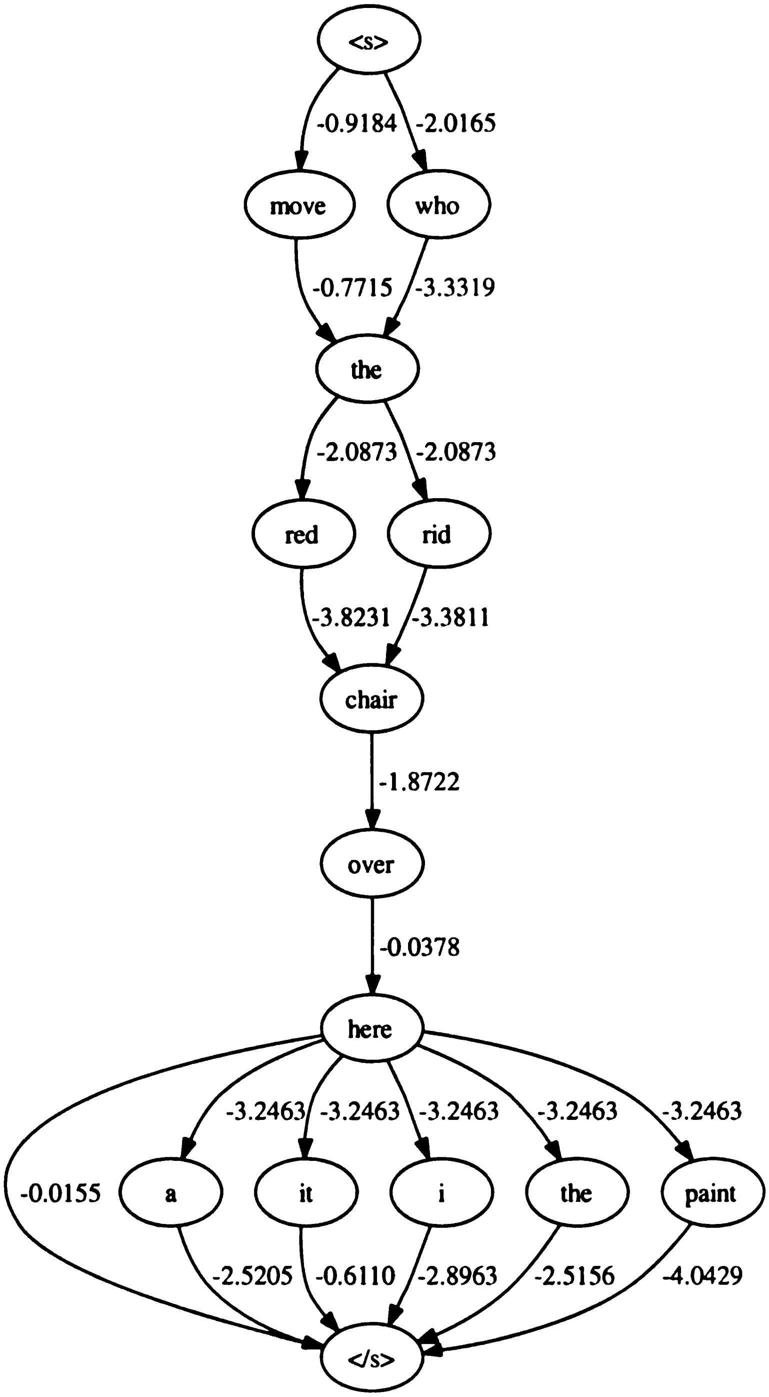


Figure 4.16. Word lattice of utterance "*move the red chair over here*" generated by using **standard bigram model**

(-3.3811) is higher than the probability of "red chair" (-3.8231). This makes the wrong speech hypothesis the top one in the n-best list (Figure 4.15). After tailoring

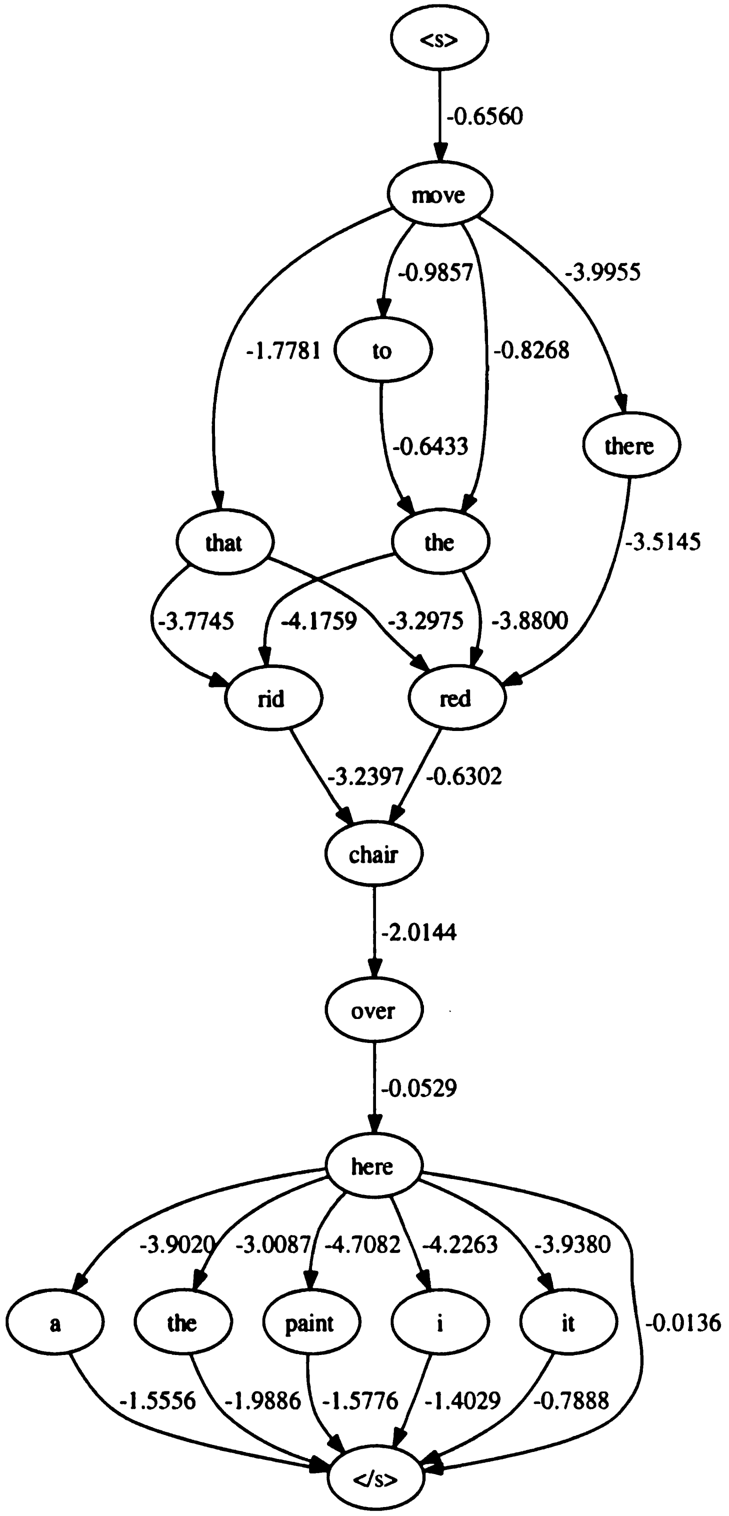


Figure 4.17. Word lattice of utterance "*move the red chair over here*" generated by using **salience driven bigram model**

the bigram model with gesture selection, in the salience driven bigram model (Figure 4.17), the probability of bigram "red chair" is much higher than the probability of "rid chair", which makes the correct speech hypothesis the best one in the n-best list and thus gets correct speech recognition.

Utterance: "I like the picture with like a forest in it"

Gaze salience:
p(bedroom) = 0.5960 p(chandelier_1) = 0.4040

Bigram n-best list:
and i eight that picture rid like got five
and i eight that picture rid identifiable
and i eight that picture rid like got forest
and i eight that picture rid like got front
and i eight that picture rid like got forest a
...

S-Bigram2 n-best list:
and i that bedroom it like upside
and i that bedroom it like a five
and i that bedroom it like a forest
and i that bedroom it like a forest a
and i that bedroom it like a forest candle
...

Figure 4.18. N-best lists of speech recognition for utterance "*I like the picture with like a forest in it*"

Unlike the active input mode of deictic gesture, eye gaze is a passive input mode. The salience information indicated by eye gaze is not as reliable as the one indicated by deictic gesture. When the salient entities indicated by eye gaze are not the true entities the user is referring to, the salience driven language model can worsen speech recognition. Figure 4.18 shows an example where the S-Bigram2 model in early application worsened the recognition of a user's utterance "*I like the picture with like a forest in it*" because of wrong salience information. In this example, the user was talking about a picture entity *picture_bamboo*. However, this entity was not salient, only entities *bedroom* and *chandelier_1* were salient as indicated by the user's eye gaze. As a result, the recognition with the S-Bigram2 model becomes worse than the baseline. The correct word "picture" is missing and the wrong word "bedroom" appears in the result.

The failure to identify the actual referred entity *picture_bamboo* as salient in the above example can also be caused by the visual properties of entities. Smaller entities on the screen are harder to be fixated by eye gaze than larger entities. To address this issue, more reliable salience modeling that takes into account the visual features is needed.

Utterance: "remove this lamp"

Gesture salience:
p(bedroom) = 0.0995
p(lamp_bank) = 0.5288
p(table_dresser) = 0.3604
p(table_pc) = 0.0114

N-best list of standard trigram model:
remove this stand
remove this them
remove this left

N-best list of S-Bigram model in early integration:
remove this lamp
remove this lamp a

N-best list of S-Bigram model in late integration:
remove this left
remove this stand
remove this them

Figure 4.19. N-best lists of an utterance: early stage integration v.s. late stage integration

Early application has an advantage over the late application on bringing the *good* hypothesized words with low acoustic probabilities into the word lattice. This is particularly important when using the Sphinx-4 speech recognizer, because the current release of Sphinx-4 does not provide a full word lattice. When the correct words are not in the word lattice output, a late application of salience driven language models will never succeed in retrieving those correct words by rescoring the word lattice.

Figure 4.19 shows one example that demonstrates the difference between the early application and the late application. Here the correct word "lamp" did not appear in the word lattice generated by the trigam model, and thus could not be retrieved by the late application of the salience driven bigram model. When the salience driven bigram model was applied in an early stage, the top one in the generated n-best list turned out to be the correct recognition result.

4.7 Summary

This chapter presents a systematic investigation of incorporating gesture/gaze into speech recognition and understanding via salience driven language modeling. Three salience driven language models based on the bigram model, the class-based bigram model, and the PCFG are compared. Our experimental results have shown that the salience driven bigram model can improve spoken language understanding in both early and late applications, while the salience driven class-based bigram model seems only useful for the late application. In the early application, the salience driven PCFG model has also shown a potential advantage in improving spoken language understanding.

CHAPTER 5

Incorporation of Non-verbal Modalities in Intention Recognition for Spoken Language Understanding

In multimodal interpretation, the user's speech is first converted to text by speech recognition. To understand the user's speech, the system further extracts semantic meaning from the user's recognized utterance. The previous chapter has addressed speech recognition in multimodal conversation. In this chapter, we address the understanding of the recognized speech during multimodal conversation.

In speech and deictic gesture systems, deictic gestures have been mainly used for attention identification (i.e., identifying which object the user is talking about). Many approaches have been developed to incorporate gestural information to resolve referring expressions (e.g., using gesture information to resolve what *this* refers to in the utterance "*how much does this cost?*") [12,14,42,54,72,119]. Different from these earlier works, our work focuses on how to take gesture beyond attention identification to help intention recognition (i.e., inferring what the user intends to do with an object), which is the main task of language understanding.

Traditional language understanding is solely based on the text input. In multimodal conversational systems, besides the user's language, it is possible to infer the context of the user's language from other non-verbal modalities (e.g. gesture) and use this context for language understanding. In speech and deictic gesture systems, deictic gestures on the graphical display indicate the user's attention, which constitutes the context of the user's utterance. Since the context of the identified attention can potentially constrain the associated intention, the deictic gestures can go beyond attention and apply to recognize the user's intention.

Within the context of a speech and gesture system, this chapter systematically investigates the role of deictic gestures in incorporating contextual information to help language understanding, specifically, to help recognize the user's intention. We experiment with different model-based and instance-based approaches to incorporate gestural information for intention recognition. We also examine the effects of using gestural information for intention recognition in two different processing stages: speech recognition stage and language understanding stage.

5.1 Multimodal Interpretation in a Speech-Gesture System

Multimodal interpretation involves extraction of semantic meanings from multimodal inputs. In human-machine conversation, the specific task of multimodal interpretation is to convert the user's multimodal input into a semantic representation that is recognizable to the system.

5.1.1 Semantic Representation

Semantic meanings from user input can be generally categorized into *intention* and *attention* [33]. Intention indicates the user's motivation and action. Attention reflects the focus of the conversation. Structuring semantic meanings in this way, we represent the semantic meaning of a user's input by a semantic frame containing intention and attention of the user. Figure 5.1 shows the semantic frame of a user's multimodal input. In the example, the user asks "*who is the artist of this picture?*" while pointing to a picture object (identified as *picture_lotus*) on the screen. The intention indicates that the user wants the artist information, whereas attention indicates *picture_lotus* is the object that the user is interested in.

Intention
action: *ACT-INFO_REQUEST*
aspect: *ARTIST*
Attention
object id: *picture_lotus*

Figure 5.1. Semantic frame of a user's multimodal input

Representing semantic meaning as semantic frames, the specific task of multimodal interpretation is to fill intention and attention units in the semantic frames based on the user's multimodal input.

5.1.2 Incorporating Context in Two Stages

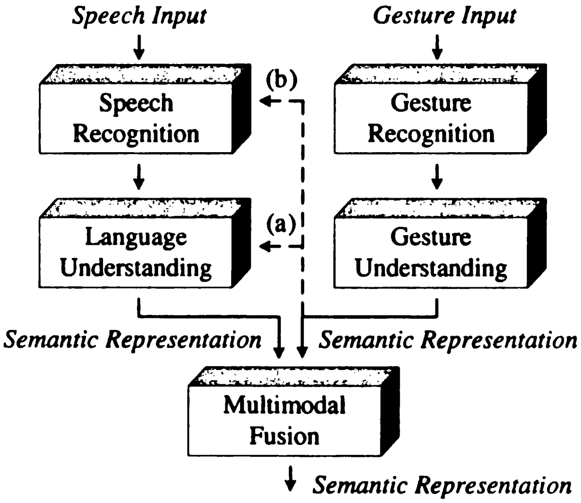


Figure 5.2. Using context (via gesture) for language understanding

Context can be incorporated in two stages to help language understanding in multimodal interpretation [83]. Take speech and gesture systems for example, as illustrated by (a) in Figure 5.2, contextual information (inferred from gesture) can be used together with recognized speech hypotheses directly in language understanding (LU) stage to improve language understanding. Since speech recognition is not perfect, and better speech recognition should lead to better language understanding, contextual information can also be used in speech recognition (SR) stage to improve speech recognition hypotheses and thus improve language understanding (Figure 5.2-(b)).

5.2 Intention Recognition

We investigate using the context identified by gesture for intention recognition in a speech-gesture system that is built for a 3D interior decoration domain (Section 3.3.1). In this domain, the user's intention is represented by an action and its corresponding aspect. All actions and corresponding aspects in the interior decoration domain are shown in Table 5.1. Note that for action *ACT-INFO_REQUEST*, the aspect includes different domain properties such as *ARTIST*, *AGE*, and *PRICE*.

Action	Aspect
ACT-ADD	<null>
ACT-ALTERNATES_SHOW	<null>
ACT-INFO_REQUEST	<domain property> or <null>
ACT-MOVE	<location> or <null>
ACT-PAINT	<color> or <null>
ACT-REMOVE	<null>
ACT-REPLACE	<replacement> or <null>
ACT-ROTATE	<direction> or <null>

Table 5.1. Intentions in the 3D interior decoration domain

Given this representation, intention recognition can be formulated as a classification problem. Each action-aspect pair can be considered as a particular type of intention. For action *ACT-INFO_REQUEST*, there are 11 possible aspect values which result in 11 classes. For all other 7 actions, each action is treated as one type of intention despite multiple possible aspect values. During interpretation, additional post-processing will take place to identify different aspects. For example, for action *ACT-PAINT*, the system will try to identify the <color> value (e.g., *red*, *blue*) from the user's utterance after *ACT-PAINT* is predicted as the user's intended action. Here, we only focus on the classification of intention without elaborating on the post-processing. In total, there are 19 target classes for intention recognition (including class *NOT-UNDERSTOOD* to represent the intention that is not supported in the domain).

5.3 Feature Extraction

To predict user intention, we first need to extract features from the user's multimodal input. Two types of features are used for intention prediction: semantic features and phoneme features.

5.3.1 Semantic Features

The semantic features of users' multimodal input consist of two parts: lexical features extracted from users' spoken utterances, and contextual features extracted from users' deictic gestures.

- Lexical features

  Lexical feature is represented by a binary feature vector which indicates what semantic concepts appear in the user's utterance. The semantic concepts are extracted from the recognized speech hypotheses (could be n-best hypotheses or 1-best hypothesis) based on lexical rules. Currently, we have 18 semantic concepts in the interior decoration domain with 130 lexical rules.

- Contextual features

  When a deictic gesture takes place, the selected object and its properties as defined in the domain are activated, which form the context of the user's utterance. This context constrains what the user is likely talking about. For example, the user is unlikely to ask the artist of a lamp or the wattage of a picture. Therefore, this context can be used to help predict user intention. For each gesture that accompanies the user's utterance, we choose the most likely object selected by the gesture and use the semantic type of the object as the contextual feature. There are 14 semantic types of objects in the domain.

5.3.2 Phoneme Features

Besides semantic features, we also use phoneme features of users' spoken utterances for intention prediction. For each speech recognition hypothesis of the user's utterance, we can get a phoneme sequence. Each phoneme sequence is treated as a phoneme feature.

User utterance: "*information on this*"
Phonemes: [ih n f er m ey sh ax n] [ao n] [dh ih s]

Speech recognition: "*and for mission on this*"
Phonemes: [ax n d] [f er] [m ih sh ax n] [ao n] [dh ih s]

Figure 5.3. Phonemes of an utterance

We give an example to show the potential of using phoneme features to help user intention prediction. As shown in Figure 5.3, the user's utterance is not correctly recognized and as a result, the semantic feature extracted from the recognized speech does not give any useful information about the user's intention of *ACT-INFO_REQUEST*. Therefore, using semantic features alone will fail to predict the user's intention. However, if we compare the two phoneme sequences of the true utterance and the speech recognition result, we can find that the phoneme sequences of the mis-recognized speech, [ax n d] [**f er**] [**m** ih **sh ax n**], is close to the true phoneme sequence [ih n **f er m** ey **sh ax n**]. This means that using phoneme sequence similarity can help recover the word "information", which is the key to identifying the user's intention in this utterance, and therefore can help predict the user's intention.

5.4 Model-Based Intention Recognition

Given an instance $\mathbf{x}$ that is represented by semantic features, we applied three classifiers to predict user intention.

- Naive Bayes

  The prediction c^* of instance $\mathbf{x}$ is given by

  $$c^* = \arg\max_c p(c|\mathbf{x}) = \arg\max_c p(c|x_1, x_2, \ldots, x_m) \tag{5.1}$$

  where x_i is the i-th feature of instance $\mathbf{x}$.

  Applying Bayes' theorem and assuming the features are conditionally independent given a class, we have

  $$\begin{aligned} p(c|\mathbf{x}) &= \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})} = \frac{p(x_1, x_2, \ldots, x_m|c)p(c)}{p(\mathbf{x})} \\ &= \frac{p(c)\prod_{i=1}^{m} p(x_i|c)}{p(\mathbf{x})} \\ &\propto p(c)\prod_{i=1}^{m} p(x_i|c) \end{aligned} \tag{5.2}$$

  Estimating $p(c)$ and $p(x_i|c)$ from the training data, we can get the prediction of a testing instance by Equation (5.1). In our evaluation, add-one smoothing was used in the estimation of $p(c)$ and $p(x_i|c)$ for predicting user intention.

- Decision Tree

  In a decision tree, each root node provides the classification of the instances, each non-leaf node specifies a test of some attribute of the instances, and each branch descending from that node corresponds to one of the possible values for this attribute. Decision trees classify instances by sorting them down the tree from the root node to some leaf node through a list of attribute tests. We used C4.5 algorithm [86] to construct decision trees for intention prediction based on the semantic features of users' multimodal input.

- Support Vector Machines (SVM)

  The SVM [22] is built by mapping instances to a high dimensional space and finding a hyperplane with the largest margin that separates the training instances into two classes in the mapped space. In prediction, an instance is classified depending the side of the hyperplane it lies in. A kernel function κ is used in SVM to achieve linear classification in the high dimensional space. Based on the semantic features of users' multimodal input, we used a polynomial kernel for user intention prediction.

  Since SVM can only handle binary classification, a "one-against-one" method is applied to use SVM for multi-class classification [39]. For a classification task of c classes, $c(c-1)/2$ SVMs are built for all pairs of classes and each SVM is trained on the data from the pair of two classes. In the testing phase, a test instance $\mathbf{x}$ is classified through a majority voting strategy. For each of the $c(c-1)/2$ binary classifiers built for class pair (c_i, c_j), if the classifier decides $\mathbf{x}$ belongs to the class c_i, the vote for class c_i increases by one. Otherwise, the vote for class c_j increases by one. After all binary classifiers have been used to vote for the classes, the one which wins the most votes is picked as the prediction of $\mathbf{x}$.

5.5 Instance-Based Intention Recognition

We also applied k-nearest neighbor (KNN), an instance-based approach, to predict user intention. Given a set of training instances with known intention, the KNN method (k=1) predicts the intention of a testing instance by finding the testing instance's closest match in the training instances and using the match's intention as the prediction.

We applied KNN to predict user intention based on semantic features and phoneme features. The similarity between a testing instance $\mathbf{x}^t$ and a training instance $\mathbf{x}^r$ is defined as

$$d_{sp}(\mathbf{x}^t, \mathbf{x}^r) = d_s(\mathbf{x}^t, \mathbf{x}^r) + d_p(\mathbf{x}^t, \mathbf{x}^r) \tag{5.3}$$

where $d_s(\mathbf{x}^t, \mathbf{x}^r)$ is the Hamming distance between the nominal semantic features and $d_p(\mathbf{x}^t, \mathbf{x}^r)$ is the distance between the phoneme features.

Hamming distance $d_s(\mathbf{x}^t, \mathbf{x}^r)$ is defined as:

$$d_s(\mathbf{x}^t, \mathbf{x}^r) = \sum_{k=1}^{m} (1 - \delta(x_k^t, x_k^r)) \tag{5.4}$$

where $x_k^t(x_k^r)$ is the k-th attribute in the semantic feature, and

$$\delta(x_k^t, x_k^r) = \begin{cases} 0 & x_k^t = x_k^r \\ 1 & x_k^t \neq x_k^r \end{cases}$$

Phonemes distance $d_p(\mathbf{x}^t, \mathbf{x}^r)$ is defined as follows based on different configurations:

- when n-best speech recognition hypotheses are used, and no gestural information is used:

$$d_p(\mathbf{x}^t, \mathbf{x}^r) = \min_k MED(P_k^t, P^r) \tag{5.5}$$

- when n-best speech recognition hypotheses are used, and gestural information (i.e., objects indicated by deictic gestures) is used

$$d_p(\mathbf{x}^t, \mathbf{x}^r) = \min_k MED(P_k^t, P^r) + w_e(o_t, o_r) \tag{5.6}$$

where,

MED – minimum edit distance

P_k^t – phonemes of the k-th speech recognition hypothesis of testing instance $\mathbf{x}^t$

P^r – phonemes of the speech transcript of training instance $\mathbf{x}^r$

$w_e(o_t, o_r)$ – distance between the object o_t selected by the gesture accompanying testing instance $\mathbf{x}^t$ and the object o_r selected by the gesture accompanying training instance $\mathbf{x}^r$ (0 if o_t and o_r are of the same semantic type, otherwise a non-zero constant)

5.6 Evaluation

We empirically evaluated the role of contextual information in intention recognition. We applied both model-based and instance-based approaches, and investigated the incorporation of contextual information for intention recognition in language understanding and speech recognition stages.

5.6.1 Experiment Settings

The CMU Sphinx-4 speech recognizer [111] was used for speech recognition. An open acoustic model and a domain dictionary were used in recognizing users' spoken utterances.

For model-based intention prediction, we evaluated the intention prediction accuracies with the following classifiers based on semantic features:

- *NBayes* – naive bayes
- *DTree* – decision tree (C4.5)
- *SVM* – support vector machine (polynomial kernel)

For instance-based intention prediction, we evaluated the intention prediction accuracies with KNN classifiers based on different instance similarity functions:

- *S-KNN* – instance distance defined on semantic features (Equation (5.4))
- *P-KNN* – instance distance defined on phoneme features (Equations (5.5) and (5.6) depending on whether gestural information is incorporated)
- *SP-KNN* – instance distance defined on combinational features of semantics and phonemes (Equation (5.3))

For each approach, we compared the performances of using only the 1-best speech recognition hypothesis and using all n-best speech recognition hypotheses for intention prediction. Also, to compare the influences of gestural information on intention prediction, we evaluated intention prediction under three gesture configurations:

- *noGest* – no gestural information is used.
- *recoGest* – with gesture recognition results, i.e., the most likely objects selected by the user's gestures as recognized by the system.
- *trueGest* – with ground truth gesture recognition results, i.e., the objects truly selected by the user's gestures.

For each approach, we further evaluated intention prediction based on standard speech recognition and gesture-tailored speech recognition. When intention prediction is based on standard speech recognition, gestural information is incorporated only in language understanding for intention prediction. When intention prediction is based on gesture-tailored speech recognition, gestural information is already used in speech recognition and can also be used in language understanding stage for intention prediction.

The evaluations were done by a 10-fold cross validation on the speech and gesture data set as described in Section 4.6.1.

5.6.2 Results Based on Traditional Speech Recognition

Table 5.2 shows the intention prediction accuracies based on the standard speech recognition results that did not use gestural information. The intention prediction accuracies based on transcripts of users' spoken utterances are also given in the table to show the upper-bound performance when speech is perfectly recognized.

		NBayes	DTree	SVM	S-KNN	P-KNN	SP-KNN
transcript	noGest	0.860	0.881	0.878	0.881	0.918	0.937
	recoGest	0.878	0.888	0.884	0.888	0.921	0.934
	trueGest	0.874	0.889	0.884	0.884	0.921	0.934
n-best hypotheses	noGest	0.709	0.718	0.713	0.700	0.790	0.824
	recoGest	0.741	0.729	0.749	0.740	0.797	0.826
	trueGest	0.755	0.738	0.744	0.737	0.806	0.832
1-best hypothesis	noGest	0.721	0.727	0.730	0.730	0.798	0.820
	recoGest	0.747	0.755	0.747	0.757	0.801	0.834
	trueGest	0.763	0.769	0.760	0.758	0.804	0.844

Table 5.2. Accuracies of intention prediction based on standard speech recognition

For all model-based approaches (i.e., NBayes, DTree, SVM), the results show that using gestural information together with recognized speech (1-best or n-best) in intention prediction achieves significant improvement on prediction accuracy compared to not using gestural information. Among instance-based approaches (i.e., S-KNN, P-KNN, SP-KNN), only for the S-KNN that uses semantic features, intention prediction accuracies are improved significantly when gestural information is used together with recognized speech (1-best or n-best hypotheses). For the P-KNN, where only phoneme features are used, there is no significant change between the intention prediction using gesture and not using gesture, no matter gestural information is used together with 1-best speech recognition or n-best speech recognition. For the SP-KNN that uses both semantic and phoneme features, intention prediction is significantly improved only when gestural information is used together with 1-best speech recognition.

It is found that, used together with recognized speech hypotheses in model-based approaches, ground truth gesture selection achieves more accurate intention prediction than recognized gesture selection in most configurations. This indicates that improving gesture recognition and understanding can further enhance intention prediction when speech recognition is not perfect. When SVM is applied on semantic features extracted from all n-best speech recognition hypotheses, using the true gesture selection achieves slightly worse performance than using the recognized gesture selection. However, this is not a significant difference. In instance-based approaches, using true gesture selection makes no significant difference than using recognized gesture selection for user intention prediction.

5.6.3 Results Based on Gesture-Tailored Speech Recognition

Table 5.3 shows the intention prediction accuracies based on the gesture-tailored speech recognition hypotheses. Note that in Table 5.3, gestural information (all possible gesture selections recognized by the system) has been utilized in speech recognition [81], the configurations *noGest*, *recoGest*, and *trueGest* only apply to how gestural information is used in language understanding stage for intention prediction. Therefore, in Table 5.3, the results under configurations *n-best hypotheses* + *noGest* and *1-best hypothesis* + *noGest* are actually the intention prediction performance when gestural information is used in only speech recognition stage.

		NBayes	DTree	SVM	S-KNN	P-KNN	SP-KNN
transcript	noGest	0.860	0.881	0.878	0.881	0.918	0.937
	recoGest	0.878	0.888	0.884	0.888	0.921	0.934
	trueGest	0.874	0.889	0.884	0.884	0.921	0.934
n-best hypotheses	noGest	0.727	0.749	0.750	0.753	0.826	0.858
	recoGest	0.753	0.766	0.780	0.770	0.829	0.857
	trueGest	0.766	0.781	0.786	0.781	0.827	0.860
1-best hypothesis	noGest	0.735	0.743	0.752	0.758	0.812	0.843
	recoGest	0.764	0.772	0.764	0.778	0.815	0.855
	trueGest	0.783	0.795	0.777	0.795	0.817	0.860

Table 5.3. Accuracies of intention prediction based on gesture-tailored speech recognition

Compared to using gestural information only in speech recognition, the accuracies of intention prediction are significantly improved in all model-based approaches when gestural information is used in both speech recognition and language understanding, no matter it is used together with 1-best or n-best speech recognition. Among instance-based approaches, only in S-KNN, that using gestural information in both speech recognition and language understanding (with 1-best or n-best recognition hypotheses) significantly improves intention prediction compared to using gestural information only in speech recognition. For P-KNN, whether or not using gestural information in language understanding does not make significant change on intention prediction. For SP-KNN, only when gestural information is used together with 1-best speech recognition hypothesis in language understanding that intention prediction is significantly improved compared to using gestural information only in speech recognition.

In all model-based approaches, together with recognized speech, using ground truth gesture selection in language understanding is found to improve intention prediction more than the recognized gesture selection. Again, this indicates that improving gesture recognition and understanding is helpful for intention prediction. In instance-based approaches, using true or recognized gesture selection in language understanding stage for intention prediction does not make significant differences when phoneme features are used.

5.6.4 Results Based on Different Sizes of Training Data

The empirical results have shown that using gestural information improves user intention recognition. To examine whether this improvement by using gestural information is dependent on the size of training data, we compare the accuracies of intention prediction with different sizes of training sets. The results of the approaches are shown in Figures 5.4–5.9. The semantic features and phoneme features are extracted from the 1-best speech recognition and the recognized gesture selection are used in intention prediction.

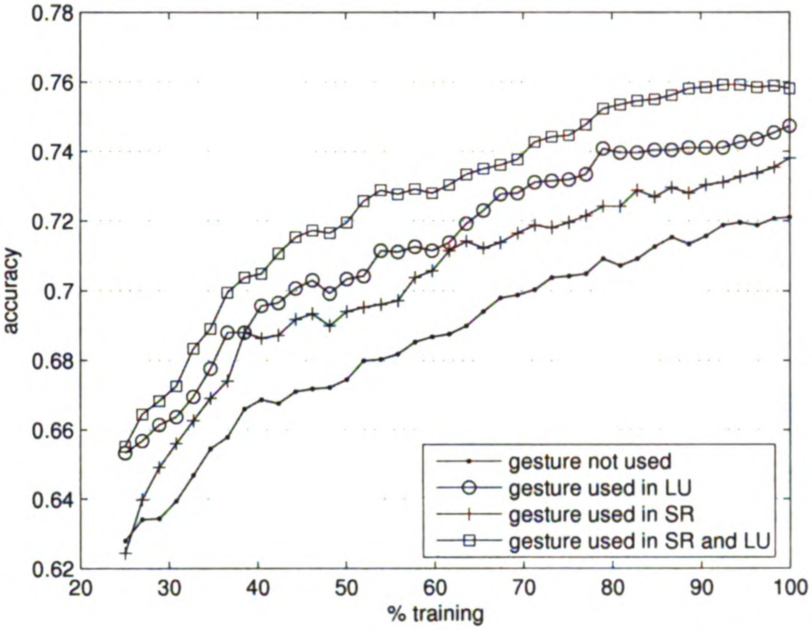


Figure 5.4. Intention prediction performance of **Naive Bayes** based on different training size

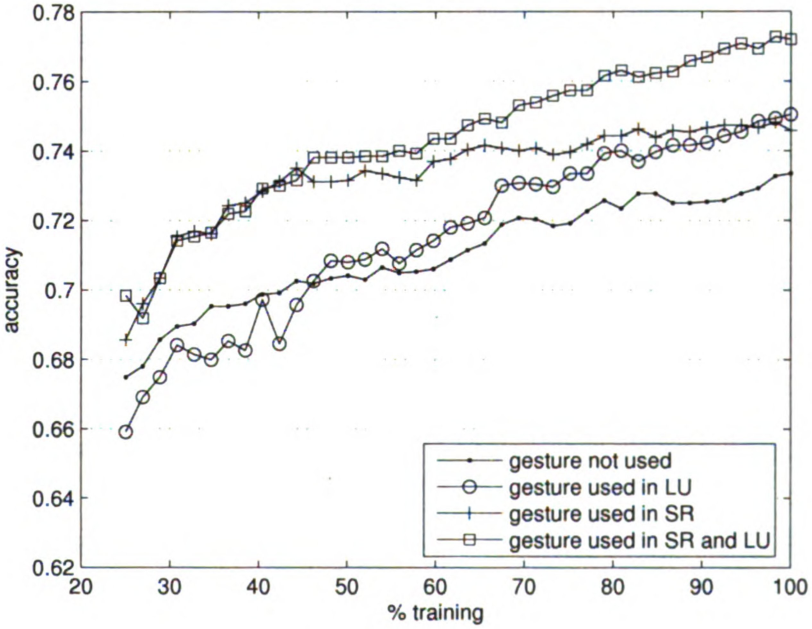


Figure 5.5. Intention prediction performance of **Decision Tree** based on different training size

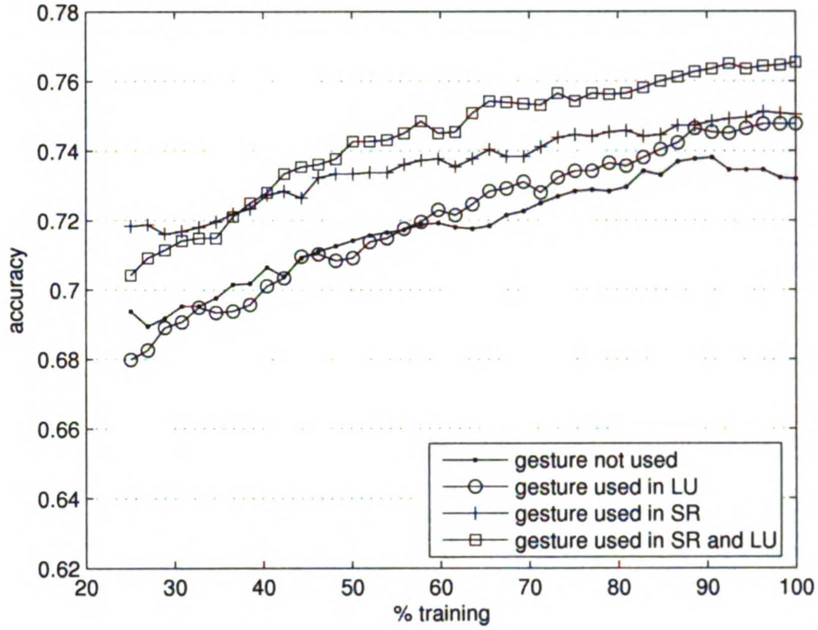


Figure 5.6. Intention prediction performance of **SVM** based on different training size

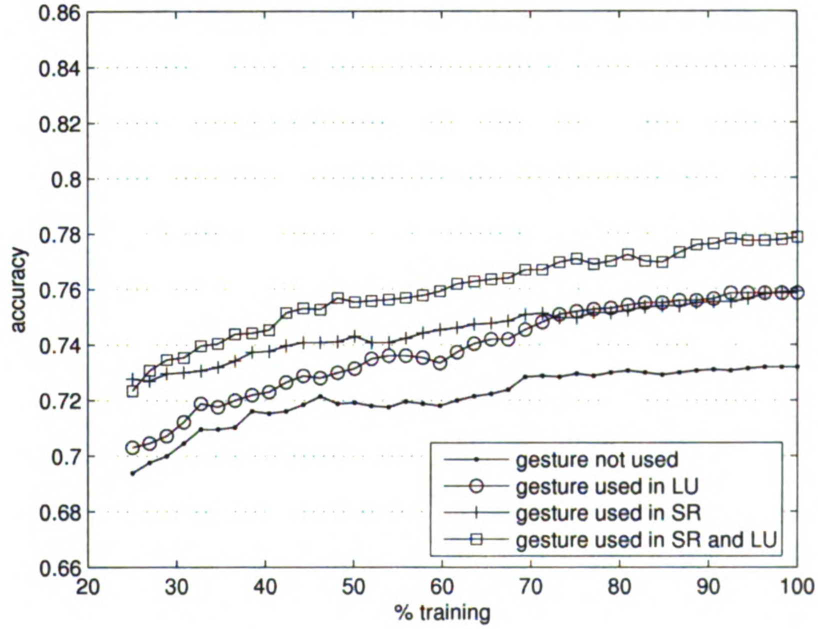


Figure 5.7. Intention prediction performance of **S-KNN** based on different training size

The intention prediction accuracy curves are generated in the following way. The whole data set is first separated into 5 folds in a stratified way such that the class

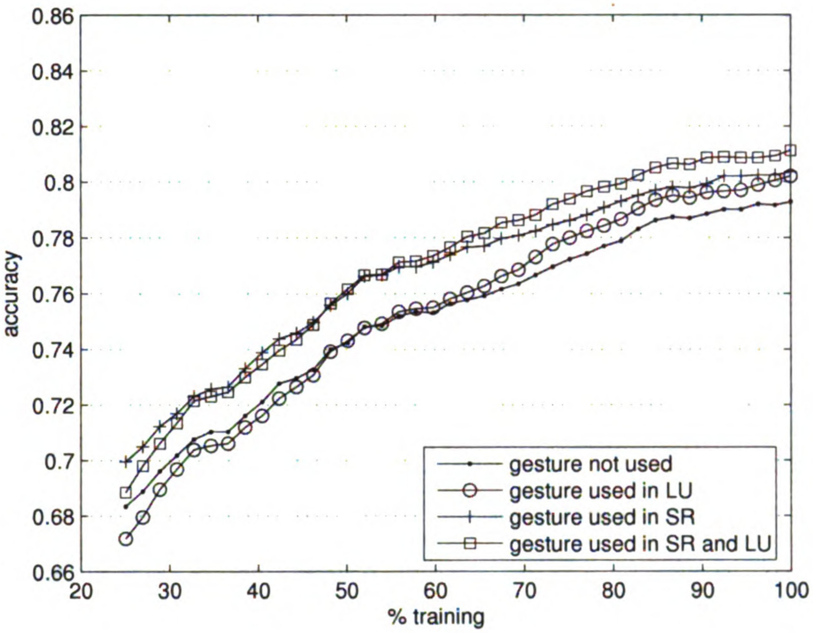


Figure 5.8. Intention prediction performance of **P-KNN** based on different training size

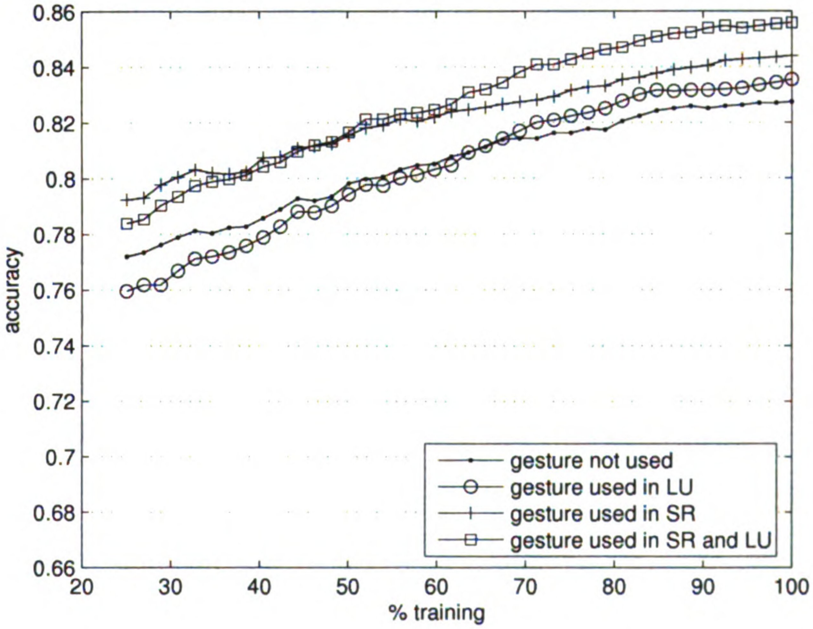


Figure 5.9. Intention prediction performance of **SP-KNN** based on different training size

distributions in each fold are the same. In each round of evaluation, two different folds are picked as the testing set and initial training set, instances in the other 3 folds are added to the training set incrementally by random picking to get intention prediction accuracies based on different sizes of training sets. After each fold of data has been used as testing set and initial training set, the intention prediction accuracy curves of the 20 round evaluations are averaged to get the curves in Figures 5.4–5.9.

We can see that, for all model-based and instance-based approaches, using gestural information in both speech recognition stage and language understanding stage always outperforms using gestural information in only language understanding stage or not using gestural information at all for intention prediction. Using gestural information only in speech recognition stage is found to always outperform not using gestural information for intention prediction in all model-based and instance-based approaches despite the training size. When gestural information is used only in language understanding stage, Naive Bayes and S-KNN always improve intention prediction despite the training size. For the other approaches (Decision Tree, SVM, P-KNN, and SP-KNN), sufficient training data is needed to make gestural information helpful for intention prediction.

5.6.5 Discussion

The empirical results lead to several findings about the role of deictic gestures in incorporating domain context in intention recognition.

First, *deictic gesture helps intention recognition given the current speech recognition technology. The earlier deictic gesture is used in the speech processing, the more effect it brings to intention recognition.* Figure 5.10 shows the performance of intention recognition by different approaches when gestural information is not used (i.e., only recognized speech hypotheses are used), used only in speech recognition stage, used only in language understanding stage, and used in both speech recognition and language understanding stages. We can easily see that using gestural information in speech recognition stage or language understanding stage improves intention prediction. Using gestural information in both speech recognition stage and language

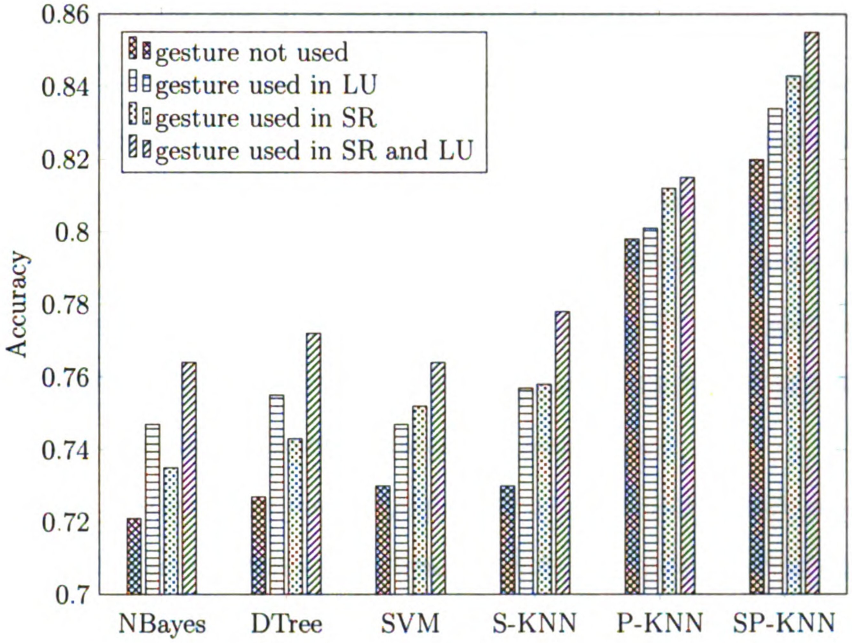


Figure 5.10. Using gestural information in different stages for intention recognition

understanding stage further improves intention prediction. Therefore, it is desirable to incorporate gesture earlier in the spoken language processing.

Second, *deictic gesture does not help much in intention recognition for a simple/small domain if speech is perfectly recognized.* As we can see in Table 5.2, when gestural information is used together with the transcripts of user utterances to predict intention, the effect is not as significant as when gesture information is used with recognized speech hypotheses. This is within our expectation. Given a simple domain with a limited number of words (the vocabulary size for our current domain is 250), it is relatively easier to come up with sufficient semantic grammars to cover the variations of language. In other words, once user utterances are correctly recognized, the semantics of the input can most likely be correctly identified by the language understanding component. So the bottleneck in interpretation appears in speech recognition (due to many possible reasons such as background noise, accent, etc.) The better is speech recognition, the better the language understanding component processes the hypotheses, and the less effect the gesture is likely to bring. When speech is perfectly recognized (i.e., same as transcriptions), the addition of gesture information will not bring extra advantage. In fact, it may hurt the performance if gesture recognition is not adequate. However, we feel that when the domain becomes more complex and the variations of language become more difficult to process, the use of gesture may begin to show advantage even when speech recognition performs reasonably well. After all, speech recognition is far from being perfect in reality, which makes gestural information valuable in intention recognition.

Third, *deictic gesture helps more significantly when combined with semantic features than with phoneme features for intention prediction.* As shown in Figure 5.10, for NBaeys, DTree, SVM and S-KNN where only semantic features are used, the addition of deictic gesture in both speech recognition and language understanding can improve the performance between 4.7% and 6.6%. For P-KNN where only the phonemes features are used, the improvement is 2.1%. Although the addition of phoneme features significantly improves the intention recognition performance, it is computationally much more expensive than the use of only semantic features. Using phoneme features may become impractical in real-time systems for complex domains. Thus the incorporation of the gestural information could be even more important.

5.7 Summary

This chapter systematically investigates the role of deictic gesture in recognizing user intention during interaction with a speech and gesture interface. Different model-based and instance-based approaches using gestural information have been applied to recognize user intention. Our empirical results have shown that using gestural information in either speech recognition or language understanding stage is able to improve user intention recognition. Moreover, when gestural information is used in both speech recognition and language understanding, intention recognition can be further improved. These results indicate that deictic gesture, although most indicative to reflect user attention, is helpful in recognizing user intention. These results further point out when and how deictic gesture should be effectively incorporated in building practical speech-gesture systems.

CHAPTER 6

Incorporation of Eye Gaze in Automatic Word Acquisition

Chapter 4 and Chapter 5 investigate the use of non-verbal modalities to improve spoken language understanding in multimodal conversational systems. Another significant problem with language understanding in multimodal conversation is the system's lack of knowledge to process user language. Language is flexible, different users may use different words to express the same meaning. When the system encounters a word that is out of its knowledge base (e.g., vocabulary), it tends to fail in interpreting the user's language. It is desirable that the system can learn new words automatically during human-machine conversation.

In this chapter, we present the investigation of using eye gaze for automatic word acquisition. The speech-gaze temporal information and domain semantic relatedness are incorporated in statistical translation models for word acquisition. Our experiments show that the use of speech-gaze temporal information and domain semantic relatedness significantly improves word acquisition performance.

This chapter begins with a description of the speech and gaze data collection, followed by an introduction of the basic translation models for word acquisition. Then, we describe the enhanced models that incorporate temporal and semantic information about speech and eye gaze for word acquisition. Finally, we present the results of empirical evaluation.

6.1 Data Collection

We used the same set of speech and eye gaze data as described in Section 4.6.3.

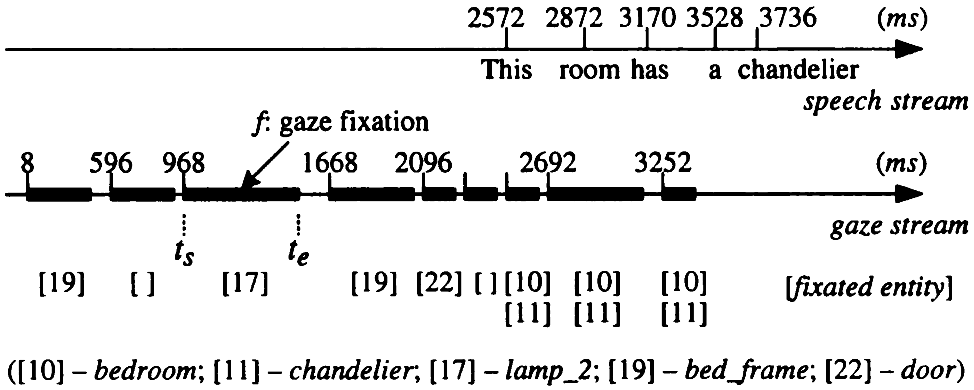


Figure 6.1. Parallel speech and gaze streams

Figure 6.1 shows an excerpt of the collected speech and gaze fixation in one experiment. In the speech stream, each word starts at a particular timestamp. In the gaze stream, each gaze fixation has a starting timestamp t_s and an ending timestamp t_e. Each gaze fixation also has a list of fixated entities (3D objects). An entity e on the graphical display is fixated by gaze fixation f if the area of e contains fixation point of f.

Given the collected speech and gaze fixations, we build a parallel speech-gaze data set as follows. For each spoken utterance and its accompanying gaze fixations, we construct a pair of word sequence and entity sequence $(\mathbf{w}, \mathbf{e})$. The word sequence $\mathbf{w}$ consists of only nouns and adjectives in the utterance. Each gaze fixation results in a fixated entity in the entity sequence $\mathbf{e}$. When multiple entities are fixated by one gaze fixation due to the overlapping of the entities, the forefront one is chosen. Also, we merge the neighboring gaze fixations that contain the same fixated entities. For the parallel speech and gaze streams shown in Figure 6.1, the resulting word sequence is $\mathbf{w}$ = [room chandelier] and the entity sequence is $\mathbf{e}$ = [*bed_frame lamp_2 bed_frame door chandelier*].

6.2 Translation Models for Automatic Word Acquisition

Since we are working on conversational systems where users interact with a visual scene, we consider the task of word acquisition as associating words with visual entities in the domain. Given the parallel speech and gaze fixated entities $\{(\mathbf{w}, \mathbf{e})\}$, we formulate word acquisition as a translation problem and use translation models to estimate word-entity association probabilities $p(w|e)$. The words with the highest association probabilities are chosen as acquired words for entity e.

6.2.1 Base Model I

Using the translation model I [5], where each word is equally likely to be aligned with each entity, we have

$$p(\mathbf{w}|\mathbf{e}) = \frac{1}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} p(w_j|e_i) \tag{6.1}$$

where l and m are the lengths of entity and word sequences respectively. We refer to this model as **Model-1**.

6.2.2 Base Model II

Using the translation model II [5], where alignments are dependent on word/entity positions and word/entity sequence lengths, we have

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p(a_j = i|j, m, l) p(w_j|e_i) \tag{6.2}$$

where $a_j = i$ means that w_j is aligned with e_i. When $a_j = 0$, w_j is not aligned with any entity (e_0 represents a *null* entity). We refer to this model as **Model-2**.

Compared to Model-1, Model-2 considers the ordering of words and entities in word acquisition. EM algorithms are used to estimate the probabilities $p(w|e)$ in the translation models.

6.3 Using Speech-Gaze Temporal Information for Word Acquisition

In Model-2, word-entity alignments are estimated from co-occurring word and entity sequences in an unsupervised way. The estimated alignments are dependent on where the words/entities appear in the word/entity sequences, not on when those words and gaze fixated entities actually occur. Motivated by the finding that users move their eyes to the mentioned object directly before speaking a word [31], we make the word-entity alignments dependent on their temporal relation in a new model (referred as **Model-2t**) [85]:

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p_t(a_j = i|j, \mathbf{e}, \mathbf{w}) p(w_j|e_i) \tag{6.3}$$

where $p_t(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the temporal alignment probability computed based on the temporal distance between entity e_i and word w_j.

We define the temporal distance between e_i and w_j as

$$d(e_i, w_j) = \begin{cases} 0 & t_s(e_i) \leq t_s(w_j) \leq t_e(e_i) \\ t_e(e_i) - t_s(w_j) & t_s(w_j) > t_e(e_i) \\ t_s(e_i) - t_s(w_j) & t_s(w_j) < t_s(e_i) \end{cases} \tag{6.4}$$

where $t_s(w_j)$ is the starting timestamp (ms) of word w_j, $t_s(e_i)$ and $t_e(e_i)$ are the starting and ending timestamps (ms) of gaze fixation on entity e_i.

The alignment of word w_j and entity e_i is decided by their temporal distance $d(e_i, w_j)$. Based on the psycholinguistic finding that eye gaze happens before a spoken word, w_j is not allowed to be aligned with e_i when w_j happens earlier than e_i (i.e., $d(e_i, w_j) > 0$). When w_j happens no earlier than e_i (i.e., $d(e_i, w_j) \leq 0$), the closer they are, the more likely they are aligned. Specifically, the temporal alignment probability of w_j and e_i in each co-occurring instance $(\mathbf{w}, \mathbf{e})$ is computed as

$$p_t(a_j = i|j, \mathbf{e}, \mathbf{w}) = \begin{cases} 0 & d(e_i, w_j) > 0 \\ \dfrac{\exp[\alpha \cdot d(e_i, w_j)]}{\sum_i \exp[\alpha \cdot d(e_i, w_j)]} & d(e_i, w_j) \leq 0 \end{cases} \tag{6.5}$$

where α is a constant for scaling $d(e_i, w_j)$.

An EM algorithm is used to estimate probabilities $p(w|e)$ and α in Model-2t.

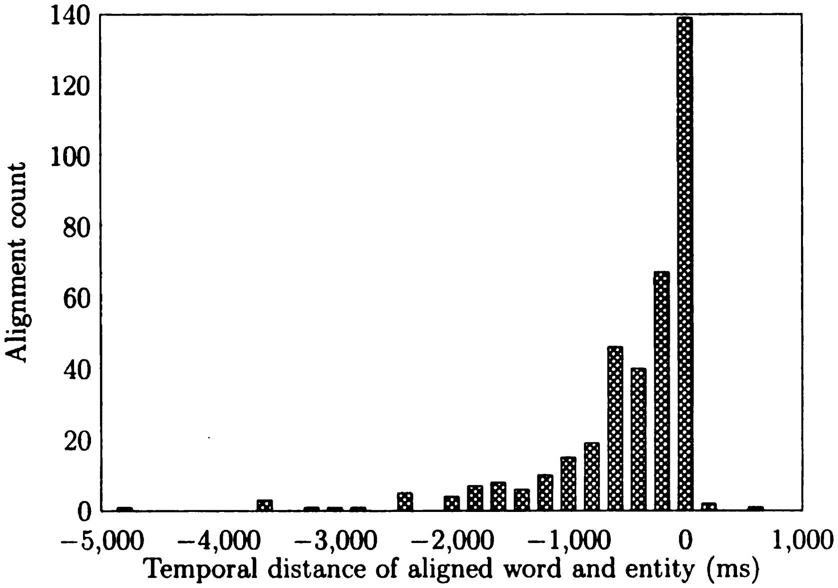


Figure 6.2. Histogram of truly aligned word and entity pairs over temporal distance (bin width = 200ms)

For the purpose of evaluation, we manually annotated the truly aligned word and entity pairs. Figure 6.2 shows the histogram of those truly aligned word and entity pairs over the temporal distance of aligned word and entity. We can observe in the figure that 1) almost no eye gaze happens after a spoken word, and 2) the number of word-entity pairs with closer temporal distance is generally larger than the number of those with farther temporal distance. This is consistent with our modeling of the temporal alignment probability of word and entity (Equation (6.5)).

6.4 Using Domain Semantic Relatedness for Word Acquisition

Speech-gaze temporal alignment and occurrence statistics sometimes are not sufficient to associate words to entities correctly. For example, suppose a user says "*there is a lamp on the dresser*" while looking at a lamp object on a table object. Due to their co-occurring with the lamp object, the words *dresser* and *lamp* are both likely to be associated with the lamp object in the translation models. As a result, the word *dresser* is likely to be incorrectly acquired for the lamp object. For the same reason, the word *lamp* could be acquired incorrectly for the table object. To solve this type of association problem, the semantic knowledge about the domain and words can be helpful. For example, the knowledge that the word *lamp* is more semantically related to the object lamp can help the system avoid associating the word *dresser* to the lamp object. Therefore, we are interested in investigating the use of semantic knowledge in word acquisition.

On one hand, each conversational system has a *domain model*, which is the knowledge representation about its domain such as the types of objects and their properties and relations. On the other hand, there are available resources about domain independent lexical knowledge (e.g., WordNet [28]). The question is whether we can use the domain model and external lexical knowledge resource to improve word acquisition. To address this question, we link the domain concepts in the domain model with WordNet concepts, and define semantic relatedness of word and entity to help the system acquire domain semantically compatible words.

In the following sections, we first describe our domain modeling, then define the semantic relatedness of word and entity based on domain modeling and WordNet semantic lexicon, and finally describe different ways of using the semantic relatedness of word and entity to help word acquisition.

6.4.1 Domain Modeling

We model the 3D room decoration domain as shown in Figure 6.3. The domain model contains all domain related semantic concepts. These concepts are linked to the WordNet concepts (i.e., synsets in the format of "word#part-of-speech#sense-id"). Each of the entities in the domain has one or more properties (e.g., semantic type, color, size) that are denoted by domain concepts. For example, the entity *dresser_1* has domain concepts *SEM_DRESSER* and *COLOR*. These domain concepts are linked to "dresser#n#4" and "color#n#1" in WordNet.

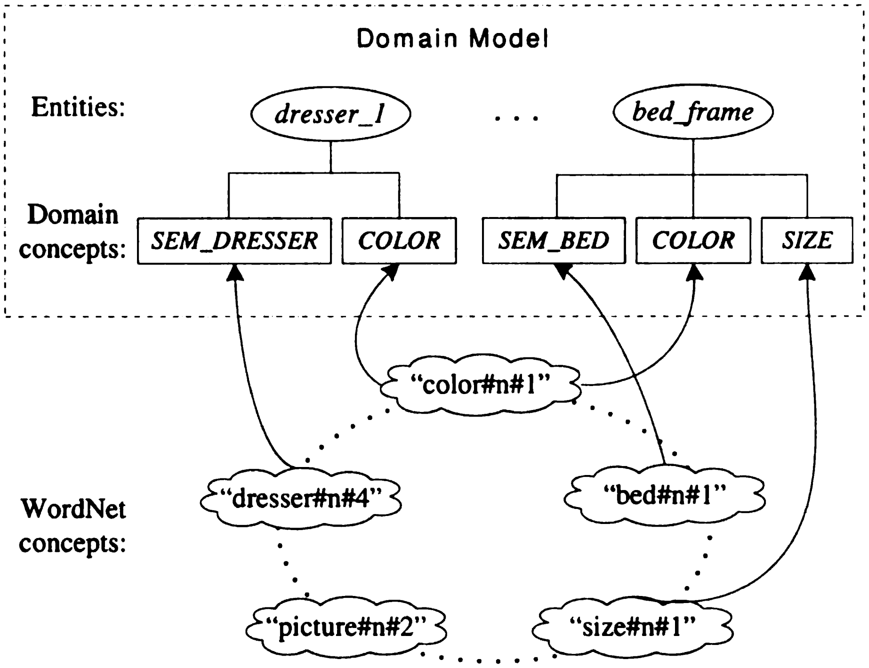


Figure 6.3. Domain model with domain concepts linked to WordNet synsets

Note that in the domain model, the domain concepts are not specific to a certain entity, they are general concepts for a certain type of entity. Multiple entities of the same type have the same properties and share the same set of domain concepts. Therefore, properties such as *color* and *size* of an entity have general concepts "color#n#1" and "size#n#1" instead of more specific concepts like "yellow#a#1" and "big#a#1", so their concepts can be shared by other entities of the same type, but with different colors and sizes.

6.4.2 Semantic Relatedness of Word and Entity

We compute the semantic relatedness of a word w and an entity e based on the semantic similarity between w and the properties of e. Specifically, semantic relatedness $SR(e,w)$ is defined as

$$SR(e,w) = \max_{i,j} sim(s(c_e^i), s_j(w)) \tag{6.6}$$

where c_e^i is the i-th property of entity e, $s(c_e^i)$ is the synset of property c_e^i as designed in domain model, $s_j(w)$ is the j-th synset of word w as defined in WordNet, and $sim(\cdot,\cdot)$ is the similarity score of two synsets.

We computed the similarity score of two synsets based on the path length between them. The similarity score is inversely proportional to the number of nodes along the shortest path between the synsets as defined in WordNet. When the two synsets are the same, they have the maximal similarity score of 1. The WordNet-Similarity tool [77] was used for the synset similarity computation.

6.4.3 Word Acquisition with Word-Entity Semantic Relatedness

We can use the semantic relatedness of word and entity to help the system acquire semantically compatible words for each entity, and therefore improve word acquisition performance. The semantic relatedness can be applied for word acquisition in two ways: post process learned word-entity association probabilities by rescoring them with semantic relatedness, or directly affect the learning of word-entity associations by constraining the alignment of word and entity in the translation models.

Rescoring with Semantic Relatedness

In the acquired word list for an entity e_i, each word w_j has an association probability $p(w_j|e_i)$ that is learned from a translation model. We use the semantic relatedness

$SR(e_i, w_j)$ to redistribute the probability mass for each w_j. The new association probability is given by:

$$p'(w_j|e_i) = \frac{p(w_j|e_i)SR(e_i, w_j)}{\sum_j p(w_j|e_i)SR(e_i, w_j)} \tag{6.7}$$

Semantic Alignment Constraint in Translation Model

When used to constrain the word-entity alignment in the translation model, semantic relatedness can be used alone or used together with speech-gaze temporal information to decide the alignment probability of word and entity [84].

- Using only semantic relatedness to constrain word-entity alignments in **Model-2s**, we have

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p_s(a_j = i|j, \mathbf{e}, \mathbf{w}) p(w_j|e_i) \tag{6.8}$$

  where $p_s(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the alignment probability based on semantic relatedness,

$$p_s(a_j = i|j, \mathbf{e}, \mathbf{w}) = \frac{SR(e_i, w_j)}{\sum_i SR(e_i, w_j)} \tag{6.9}$$

- Using semantic relatedness and temporal information to constrain word-entity alignments in **Model-2ts**, we have

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p_{ts}(a_j = i|j, \mathbf{e}, \mathbf{w}) p(w_j|e_i) \tag{6.10}$$

  where $p_{ts}(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the alignment probability that is decided by both temporal relation and semantic relatedness of e_i and w_j,

$$p_{ts}(a_j = i|j, \mathbf{e}, \mathbf{w}) = \frac{p_s(a_j = i|j, \mathbf{e}, \mathbf{w}) p_t(a_j = i|j, \mathbf{e}, \mathbf{w})}{\sum_i p_s(a_j = i|j, \mathbf{e}, \mathbf{w}) p_t(a_j = i|j, \mathbf{e}, \mathbf{w})} \tag{6.11}$$

  where $p_s(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the semantic alignment probability in Equation (6.9), and $p_t(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the temporal alignment probability given in Equation (6.5).

EM algorithms are used to estimate $p(w|e)$ in Model-2s and Model-2ts.

6.5 Grounding Words to Domain Concepts

As discussed above, based on translation models, we can incorporate temporal and domain semantic information to obtain $p(w|e)$. This probability only provides a means to ground words to entities. In conversational systems, the ultimate goal of word acquisition is to make the system understand the semantic meaning of new words. Word acquisition by grounding words to objects is not always sufficient for identifying their semantic meanings. Suppose the word *green* is grounded to a green chair object, so is the word *chair*. Although the system is aware that *green* is some word describing the green chair, it does not know that the word *green* refers to the chair's color while the word *chair* refers to the chair's semantic type. Thus, after learning the word-entity associations $p(w|e)$ by the translation models, we need to further ground words to domain concepts of entity properties.

We further apply WordNet to ground words to domain concepts. For each entity e, based on association probabilities $p(w|e)$, we can choose the n-best words as acquired words for e. Those n-best words have the n highest association probabilities. For each word w acquired for e, the grounded concept c_e^* for w is chosen as the one that has the highest semantic relatedness with w:

$$c_e^* = \arg\max_i \left[\max_j sim(s(c_e^i), s_j(w))\right] \tag{6.12}$$

where $sim(s(c_e^i), s_j(w))$ is the semantic similarity score defined in Equation (6.6).

6.6 Evaluation

To evaluate the acquired words for the entities, we manually compile a set of "gold standard" words from all users' speech transcripts and gaze fixations. Those "gold standard" words are the words that the users have used to refer to the entities and their properties (e.g., color, size, shape) during the interaction with the system. The automatically acquired words are evaluated against those "gold standard" words.

6.6.1 Evaluation Metrics

The following metrics are used to evaluate the words acquired for domain concepts (i.e., entity properties) $\{c_e^i\}$.

- Precision

$$\frac{\sum_e \sum_i \# \text{ words correctly acquired for } c_e^i}{\sum_e \sum_i \# \text{ words acquired for } c_e^i}$$

- Recall

$$\frac{\sum_e \sum_i \# \text{ words correctly acquired for } c_e^i}{\sum_e \sum_i \# \text{ ``gold standard'' words of } c_e^i}$$

- F-measure

$$\frac{2 \times precision \times recall}{precision + recall}$$

The metrics of precision, recall, and F-measure are based on the n-best words acquired for the entity properties. Therefore, we have different precision, recall, and F-measure when n changes.

The metrics of precision, recall, and F-measure only provide evaluation on the top n candidate words. To measure the acquisition performance on the entire ranked list of candidate words, we define a new metric as follows:

- Mean Reciprocal Rank Rate (MRRR)

$$\text{MRRR} = \frac{\sum_e \frac{\sum_{i=1}^{N_e} \frac{1}{index(w_e^i)}}{\sum_{i=1}^{N_e} \frac{1}{i}}}{\#e}$$

where N_e is the number of all ground-truth words $\{w_e^i\}$ for entity e, $index(w_e^i)$ is the index of word w_e^i in the ranked list of candidate words for entity e.

Entities may have a different number of ground-truth words. For each entity e, we calculate a Reciprocal Rank Rate (RRR), which measures how close the ranks of the ground-truth words in the candidate word list is to the best scenario where the top N_e words are the ground-truth words for e. RRR is in the range of $(0, 1]$. The higher the RRR, the better is the word acquisition performance. The average of RRRs across all entities gives the Mean Reciprocal Rank Rate (MRRR).

Note that MRRR is directly based on the learned word-entity associations $p(w|e)$, it is in fact a measure of grounding words to entities.

6.6.2 Evaluation Results

To compare the effects of different speech-gaze alignments on word acquisition, we evaluate the following models:

- Model-1 – base model I without word-entity alignment (Equation (6.1)).
- Model-2 – base model II with positional alignment (Equation (6.2)).
- Model-2t – enhanced model with temporal alignment (Equation (6.3)).
- Model-2s – enhanced model with semantic alignment (Equation (6.8)).
- Model-2ts – enhanced model with both temporal and semantic alignment (Equation (6.10)).

To compare the different ways of incorporating semantic relatedness in word acquisition as discussed in Section 6.4.3, we also evaluate the following models:

- Model-1-r – Model-1 with semantic relatedness rescoring of word-entity association.

- Model-2t-r – Model-2t with semantic relatedness rescoring of word-entity association.

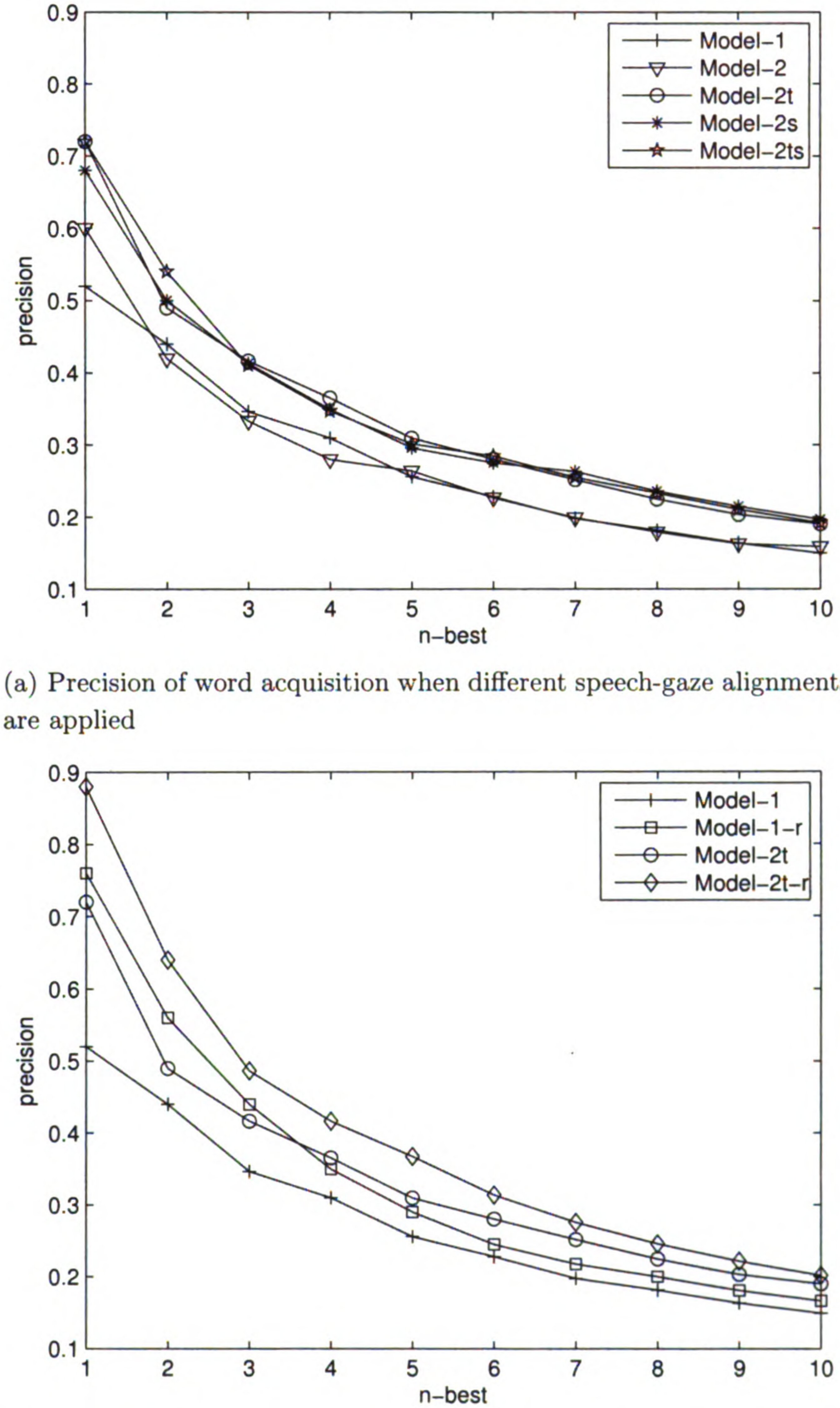
Figures 6.4, 6.5, and 6.6 compare the results of models with different speech-gaze alignments and models with semantic relatedness rescoring. In the figures, *n-best* means the top n word candidates are chosen as acquired words for each entity. The Mean Reciprocal Rank Rates of all models are compared in Figure 6.7.

Results of Using Different Speech-Gaze Alignments

As shown in Figures 6.4(a), 6.5(a), and 6.6(a), Model-2 does not show a consistent improvement compared to Model-1 when a different number of n-best words are chosen as acquired words. This result shows that it is not very helpful to consider the index-based positional alignment of word and entity for word acquisition.

Figures 6.4(a), 6.5(a), and 6.6(a) also show that models considering temporal or/and semantic information (Model-2t, Model-2s, Model-2ts) consistently perform better than the models considering neither temporal nor semantic information (Model-1, Model-2). Among Model-2t, Model-2s, and Model-2ts, it is found that they do not make consistent differences.

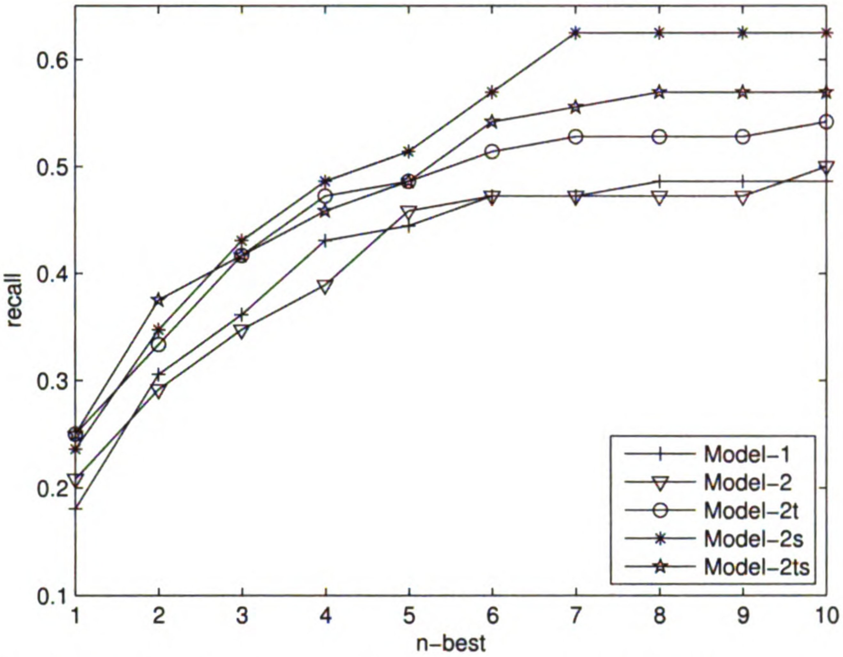
As shown in Figure 6.7, the MRRRs of different models are consistent with their performances on F-measure. A t-test has shown that the difference between the MRRRs of Model-1 and Model-2 is not statistically significant. Compared to Model-1, t-tests have confirmed that MRRR is significantly improved by Model-2t ($t = 2.29, p < 0.016$), Model-2s ($t = 3.40, p < 0.002$), and Model-2ts($t = 3.12, p < 0.003$). T-tests have shown no significant differences among Model-2t, Model-2s, and Model-2ts.



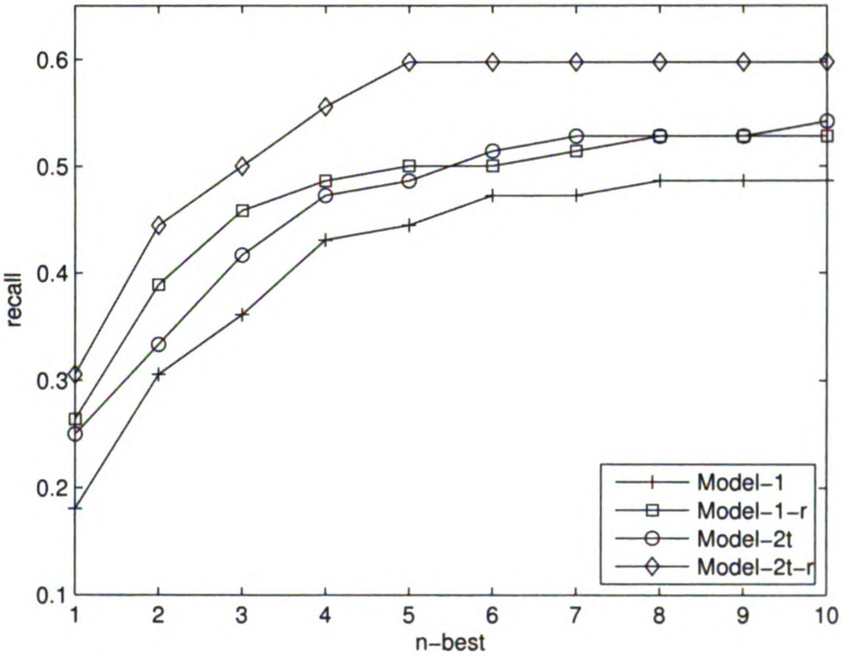
(a) Precision of word acquisition when different speech-gaze alignments are applied

(b) Precision of word acquisition when semantic relatedness rescoring of word-entity association is applied

Figure 6.4. Precision of word acquisition

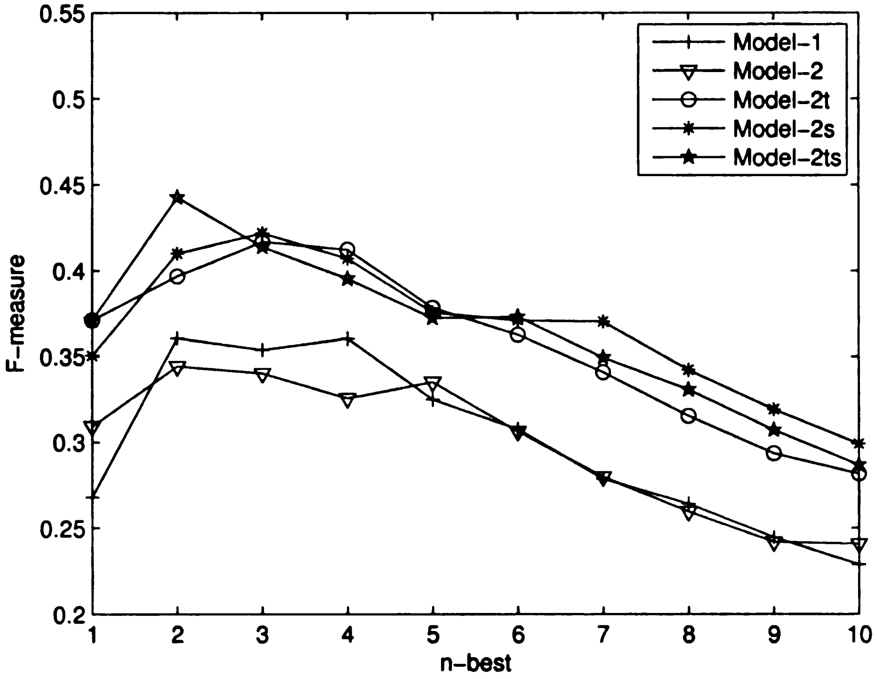


(a) Recall of word acquisition when different speech-gaze alignments are applied

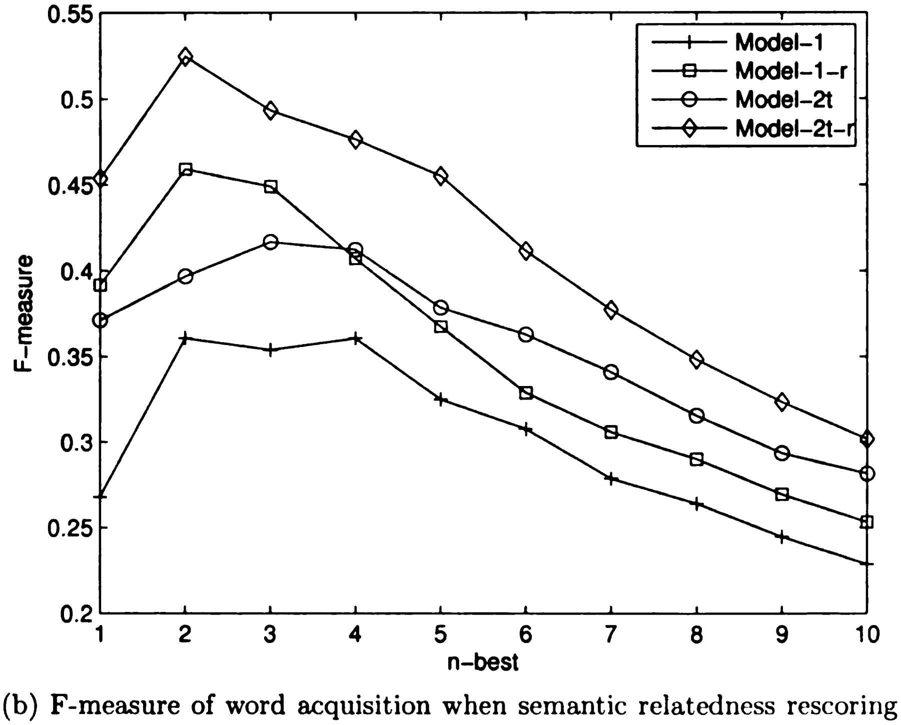


(b) Recall of word acquisition when semantic relatedness rescoring of word-entity association is applied

Figure 6.5. Recall of word acquisition



(a) F-measure of word acquisition when different speech-gaze alignments are applied



(b) F-measure of word acquisition when semantic relatedness rescoring of word-entity association is applied

Figure 6.6. F-measure of word acquisition

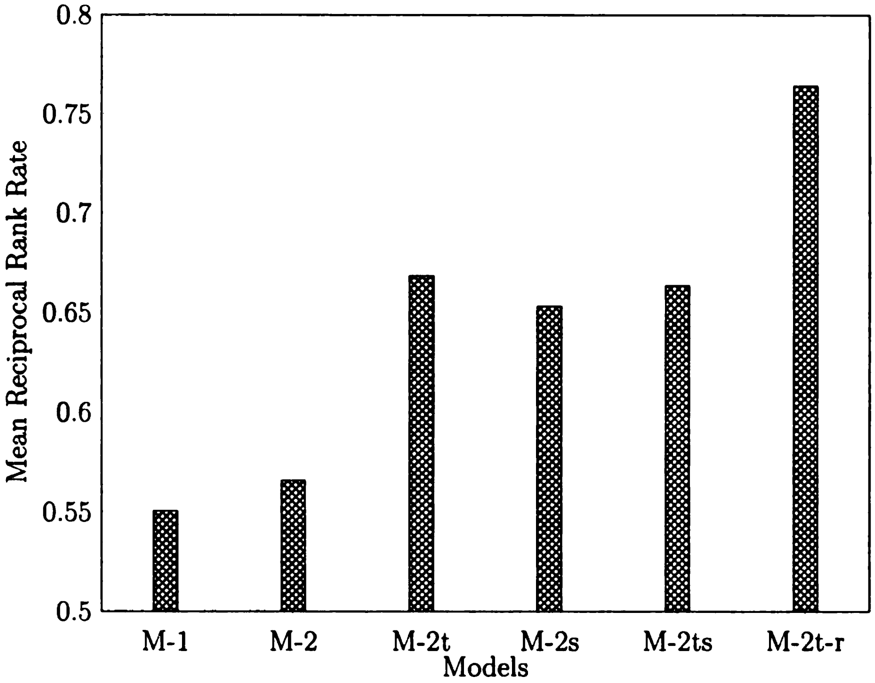


Figure 6.7. MRRRs achieved by different models

Results of Applying Semantic Relatedness Rescoring

Figures 6.4(b), 6.5(b), and 6.6(b) show that semantic relatedness rescoring improves word acquisition. After semantic relatedness rescoring of the word-entity associations learned by Model-1, Model-1-r improves the F-measure consistently when a different number of n-best words are chosen as acquired words. Compared to Model-2t, Model-2t-r also improves the F-measure consistently.

Comparing the two ways of using semantic relatedness for word acquisition, it is found that rescoring word-entity association with semantic relatedness works better. When semantic relatedness is used together with temporal information to constrain word-entity alignments in Model-2ts, word acquisition performance is not improved compared to Model-2t. However, using semantic relatedness to rescore word-entity association learned by Model-2t, Model-2t-r further improves word acquisition.

As shown in Figure 6.7, the MRRRs of Model-1-r and Model-2t-r are consistent with their performances on F-measure. Compared to Model-2t, Model-2t-r improves

MRRR. A t-test has confirmed that this is a significant improvement ($t = 1.96, p < 0.031$). Compared to Model-1, Model-1-r significantly improves MRRR ($t = 2.33, p < 0.015$). There is no significant difference between Model-1-r and Model-2t/Model-2s/Model-2ts.

In Figure 6.5, we notice that the recall of the acquired words is still comparably low even when 10 best word candidates are chosen for each entity. This is mainly due to the scarcity of those words that are not acquired in the data. Many of the words that are not acquired appear less than 3 times in the data, which makes them unlikely to be associated with any entity by the translation models. When more data is available, we expect to see higher recall.

6.6.3 An Example

Table 6.1 shows the 5-best words acquired by different models for the entity *dresser_1* in the 3D room scene. In the table, each word is followed by its word-entity association probability $p(w|e)$. The correctly acquired words are shown in bold font.

Model	Model-1	Model-2t	Model-2t-r
Rank 1	**table**(0.173)	**table**(0.196)	**table**(0.294)
Rank 2	**dresser**(0.067)	**dresser**(0.101)	**dresser**(0.291)
Rank 3	area(0.058)	area(0.056)	**vanity**(0.147)
Rank 4	picture(0.053)	**vanity**(0.051)	**desk**(0.038)
Rank 5	dressing(0.041)	dressing(0.050)	area(0.024)

Table 6.1. N-best candidate words acquired for the entity *dresser_1* by different models

As shown in the example, the baseline Model-1 learned 2 correct words in the 5-best list. Considering speech-gaze temporal information, Model-2t learned one more correct word *vanity* in the 5-best list. With semantic relatedness rescoring, Model-2t-r further acquired word *desk* in the 5-best list because of the high semantic relatedness of word *desk* and the type of entity *dresser_1*. Although neither Model-1 nor Model-2t successfully acquired the word *desk* in the 5-best list, the rank (=7) of the word *desk* in Model-2t's n-best list is much higher than the rank (=21) in Model-1's n-best list.

6.7 Summary

This chapter investigates the use of eye gaze for automatic word acquisition in multimodal conversational systems. Particularly, we investigate the use of speech-gaze temporal information and word-entity semantic relatedness to facilitate word acquisition. The experiments show that word acquisition is significantly improved when temporal information is considered, which is consistent with the previous psycholinguistic findings about speech and eye gaze. Moreover, using temporal information together with semantic relatedness rescoring further improves word acquisition.

CHAPTER 7

Incorporation of Interactivity with Eye Gaze for Automatic Word Acquisition

In the previous chapter, we describe the use of the speech-gaze temporal information and domain semantic relatedness for automatically acquiring words from the user's speech and its accompanying gaze fixations. Successful word acquisition relies on the tight link between what the user says and what the user sees. Although published studies provide us with a sound empirical basis for assuming that eye movements are predictive of speech, the gaze behavior in an interactive setting can be much more complex. There are different types of eye movements [50]. The naturally occurring eye gaze during speech production may serve different functions, for example, to engage in the conversation or to manage turn taking [70]. Furthermore, while interacting with a graphic display, a user could be talking about objects that were previously seen on the display or something completely unrelated to any object the user is looking at. Therefore using all the speech-gaze pairs for word acquisition can be detrimental. The type of gaze that is mostly useful for word acquisition is the kind that reflects the underlying attention and tightly links to the content of the co-occurring spoken utterances. Thus, one important question is how to identify the closely coupled speech and gaze streams to improve word acquisition.

To address this question, in this chapter, we develop an approach that incorporates interactivity (e.g., user activity, conversation context) with eye gaze to identify the closely coupled speech and gaze streams. We further use the identified speech and gaze streams for word acquisition. Our studies indicate that automatic identification of closely coupled gaze-speech stream pairs is an important first step that leads to performance gains in word acquisition. Our simulation studies further demonstrate the effect of automatic online word acquisition on improving language understanding in human-machine conversation.

In the following sections, we first describe the data collection in a new 3D interactive domain, then present the automatic identification of the closely coupled gaze-speech pairs and its effect on word acquisition. The last part of this chapter presents a simulation study that exemplifies how word acquisition can be automatically achieved and how the acquired words affect language interpretation during online conversation.

7.1 Data Collection

We recruited 20 users to interact with our speech-gaze system to collect data.

7.1.1 Domain

We used the 3D treasure hunting domain (see Section 3.3.2) for the investigation of automatic word acquisition in multimodal conversation. In this application, the user needs to consult with a remote "expert" (i.e., an artificial system) to find hidden treasures in a castle with 115 3D objects. The expert has some knowledge about the treasures but can not see the castle. The user has to talk to the expert for advices of finding the treasures. The application is developed based on a game engine and provides an immersive environment for the user to navigate in the 3D space. A detailed description of the user study is given in Appendix A.3.

During the experiment, the user's speech was recorded, and the user's eye gaze was captured by a Tobii eye tracker. Figure 7.1 shows a snapshot of one user's experiment.

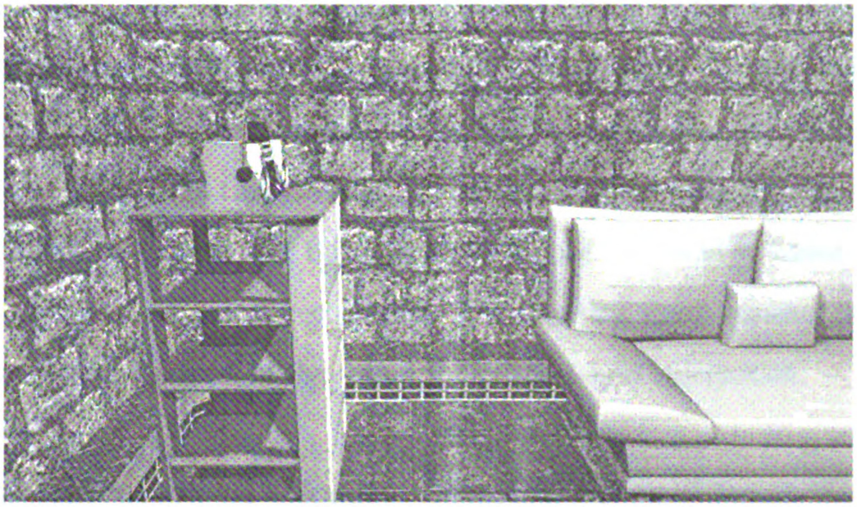


Figure 7.1. A snapshot of one user's experiment (the dot on the stereo indicates the user's gaze fixation, which was not shown to the user during the experiment)

It's worthwhile to note that the collected data set is different from the data set used for the investigation in Chapter 6. The difference lies in two aspects: 1) the data for this investigation was collected during mixed initiative human-machine conversation whereas the data in Chapter 6 was based only on question and answering; 2) user studies were conducted in a more complex domain for this investigation, which resulted in a richer data set that contains larger vocabulary.

7.1.2 Data Preprocessing

From 20 users' experiments, we collected 3709 utterances with accompanying gaze fixations. We transcribed the collected speech. The vocabulary size of the speech transcript is 1082, among which 227 words are nouns and adjectives. The user's speech was also automatically recognized online by the Microsoft speech recognizer with a word error rate (WER) of 48.1% for the 1-best recognition. The vocabulary size of the 1-best speech recognition is 3041, among which 1643 are nouns and adjectives.

The collected speech and gaze streams are automatically paired together by the system. Each time the system detects a sentence boundary of the user's speech, it pairs the recognized speech with the gaze fixations that the system has been accumulating since the previously detected sentence boundary. Given the paired speech and gaze streams, we build a parallel data set of word sequence and gaze fixated entity sequence $\{(\mathbf{w}, \mathbf{e})\}$ for the task of word acquisition. For the gaze stream, $\mathbf{e}$ contains all the gaze fixated entities. For the speech stream, we can build $\mathbf{w}$ based on speech transcript or the 1-best speech recognition. The resulting word sequence $\mathbf{w}$ contains all the nouns and adjectives in the transcript or the 1-best recognition.

7.2 Identification of Closely Coupled Gaze-Speech Pairs

As mentioned earlier, not all gaze-speech pairs are useful for word acquisition. In a gaze-speech pair, if the speech does not have any word that relates to any of the gaze fixated entities, this instance only adds noise to word acquisition. Therefore, we should identify the closely coupled gaze-speech pairs and only use them for word acquisition.

In this section, we first describe the feature extraction, then describe the use of a logistic regression classifier to predict whether a gaze-speech pair is a **closely coupled gaze-speech instance** – an instance where at least one noun or adjective in the speech stream is referring to some gaze fixated entity in the gaze stream. For the training of the classifier for gaze-speech prediction, we manually labeled each instance whether it is a closely coupled gaze-speech instance based on the speech transcript and gaze fixations.

7.2.1 Features Extraction

For a parallel gaze-speech instance, the following sets of features are automatically extracted.

Speech Features (S-Feat)

Let c_w be the count of nouns and adjectives in the utterance, and l_s be the temporal length of the speech. The following features are extracted from speech:

- c_w – count of nouns and adjectives.

  More nouns and adjectives are expected in the user's utterance describing entities.

- c_w/l_s – normalized noun/adjective count.

  The effect of speech length l_s on c_w is considered.

Gaze Features (G-Feat)

For each fixated entity e_i, let l_e^i be its fixation temporal length. Note that several gaze fixations may have the same fixated entity, l_e^i is the total length of all the gaze fixations that fixate on entity e_i. We extract the following features from gaze stream:

- c_e – count of different gaze fixated entities.

  Less fixated entities are expected when the user is describing entities while looking at them.

- c_e/l_s – normalized entity count.

  The effect of speech temporal length l_s on c_e is considered.

- $\max_i(l_e^i)$ – maximal fixation length.

  At least one fixated entity's fixation is expected to be long enough when the user is describing entities while looking at them.

- $mean(l_e^i)$ – average fixation length.

  The average gaze fixation length is expected to be longer when the user is describing entities while looking at them.

- *$var(l_e^i)$* – variance of fixation lengths.

  The variance of the fixation lengths is expected to be smaller when the user is describing entities while looking at them.

The number of gaze fixated entities is not only decided by the user's eye gaze, it is also affected by the visual scene. Let c_e^s be the count of all the entities that have been visible during the length of the gaze stream. We also extract the following scene related feature:

- c_e/c_e^s – scene normalized fixated entity count.

  The effect of the visual scene on c_e is considered.

User Activity Features (UA-Feat)

While interacting with the system, the user's activity can also be helpful in determining whether the user's eye gaze is tightly linked to the content of the speech. The following features are extracted from the user's activities:

- *maximal distance of the user's movements* – the maximal change of user position (3D coordinates) during the speech length.

  The user is expected to move within a smaller range while looking at entities and describing them.

- *variance of the user's positions*

  The user is expected to move less frequently while looking at entities and describing them.

Conversation Context Features (CC-Feat)

While talking to the system (i.e., the "expert"), the user's language and gaze behavior are influenced by the state of the conversation. For each gaze-speech instance, we use the previous system response type as a nominal feature to predict whether this is a closely coupled gaze-speech instance.

In our treasure hunting domain, there are 8 types of system responses in 2 categories:

System Initiative Responses:

- *specific-see* – the system asks whether the user sees a certain entity, e.g., "Do you see another couch?".
- *nonspecific-see* – the system asks whether the user sees anything, e.g., "Do you see anything else?", "Tell me what you see".
- *previous-see* – the system asks whether the user previously sees something, e.g., "Have you previously seen a similar object?".
- *describe* – the system asks the user to describe in detail what the user sees, e.g., "Describe it", "Tell me more about it".
- *compare* – the system asks the user to compare what the user sees, e.g., "Compare these objects".
- *clarify* – the system asks the user to make clarification, e.g., "I did not understand that", "Please repeat that".
- *action-request* – the system asks the user to take action, e.g., "Go back", "Try moving it".

User Initiative Responses:

- *misc* – the system hands the initiative back to the user without specifying further requirements, e.g., "I don't know", "Yes".

7.2.2 Logistic Regression Model

Given the extracted feature $\mathbf{x}$ and the "closely coupled" label y of each instance in the training set, we train a ridge logistic regression model [60] to predict whether an instance is a closely coupled instance ($y = 1$) or not ($y = 0$).

In the logistic regression model, the probability that $y_i = 1$, given the feature $\mathbf{x}_i = (x_1^i, x_2^i, \ldots, x_m^i)$, is modeled by

$$p(y_i|\mathbf{x}_i) = \frac{\exp(\sum_{j=1}^{m} \beta_j x_j^i)}{1 + \exp(\sum_{j=1}^{m} \beta_j x_j^i)}$$

where β_j are the feature's weights to be learned.

The log-likelihood l of the data $(\mathbf{X}, \mathbf{y})$ is

$$l(\beta) = \sum_i [y_i \log p(y_i|\mathbf{x}_i) + (1 - y_i) \log(1 - p(y_i|\mathbf{x_i}))]$$

In ridge logistic regression, parameters β_j are estimated by maximizing a regularized log-likelihood

$$l^\lambda(\beta) = l(\beta) - \lambda||\beta||^2$$

where λ is the ridge parameter that is introduced to achieve more stable parameter estimation.

We used the Weka toolkit [115] for the training of the ridge logistic regression model.

7.3 Evaluation of Gaze-Speech Identification

We evaluate the gaze-speech identification for the instances with 1-best speech recognition. Since the goal of identifying closely coupled gaze-speech instances is to improve word acquisition and we are only interested in acquiring nouns and adjectives, only the instances with recognized nouns/adjectives are used for training the logistic regression classifier. Among the 2969 instances with recognized nouns/adjectives and gaze fixations, 2002 (67.4%) instances are labeled as closely coupled. The gaze-speech prediction was evaluated by a 10-fold cross validation.

Table 7.1 shows the prediction precision and recall when different sets of features are used. As seen in the table, as more features are used, the prediction precision goes up and the recall goes down. It is important to note that prediction precision is more critical than recall for word acquisition when sufficient amount data is available. *Noisy* instances where the gaze does not link to the speech content will only hurt word acquisition since they will guide the translation models to ground words to the wrong entities. Although higher recall can be helpful, its effect is expected to become less when more data becomes available.

Feature sets	Precision	Recall
Null (*baseline*)	0.674	1
S-Feat	0.686	0.995
G-Feat	0.707	0.958
UA-Feat	0.704	0.942
CC-Feat	0.688	0.936
G-Feat + UA-Feat	0.719	0.948
G-Feat + UA-Feat + S-Feat	0.741	0.908
G-Feat + UA-Feat + CC-Feat	0.731	0.918
G-Feat + UA-Feat + S-Feat + CC-Feat	**0.748**	0.899

Table 7.1. Gaze-speech prediction performances with different feature sets for the instances with 1-best speech recognition

The results show that speech features (S-Feat) and conversation context features (CC-Feat), when used alone, do not improve prediction precision much compared to the baseline of predicting all instances "closely coupled" with a precision of 67.4%. When used alone, gaze features (G-Feat) and user activity features (UA-Feat) are the two most useful feature sets for increasing prediction precision. When they are used together, the prediction precision is further increased. Adding either speech features or conversation context features to gaze and user activity features (G-Feat + UA-Feat + S-Feat/CC-Feat) increases the prediction precision more. Using all four sets of features (G-Feat + UA-Feat + S-Feat + CC-Feat) achieves the highest prediction precision, which is significantly better than the baseline: $z = 5.93, p < 0.001$. Therefore, we choose to use all feature sets to identify the closely coupled gaze-speech instances for word acquisition.

To compare the effect of the identified closely coupled gaze-speech instances on word acquisition from different speech input (1-best speech recognition, speech transcript), we also use the logistic regression classifier with all features to predict closely coupled gaze-speech instances for the instances with speech transcript. For the instances with speech transcript, there are 2948 instances with nouns/adjectives and gaze fixations, 2128 (72.2%) of them being labeled as closely coupled. The prediction precision is 77.9% and the recall is 93.8%. The prediction precision is significantly better than the baseline of predicting all instances as coupled: $z = 4.92, p < 0.001$.

7.4 Evaluation of Word Acquisition

In Chapter 6, we have shown that Model-2t-r (Section 6.4), where the temporal alignment between speech and eye gaze and domain semantic relatedness are incorporated, achieves significantly better word acquisition performance. Therefore, this model is used for the word acquisition in this investigation. The word acquired by Model-2t-r are evaluated against the "gold standard" words that we manually compiled for each entity and its properties based on all users' speech transcripts and gaze fixations. Those "gold standard" words are the words that the users have used to describe the entities and their properties during the interaction with the system.

7.4.1 Evaluation Metrics

We evaluate the n-best acquired words on

- Precision

- Recall

- F-measure

When a differen n is chosen, we will have different precision, recall, and F-measure. We also evaluate the whole ranked candidate word list on

- Mean Reciprocal Rank Rate (MRRR) (see Section 6.6.1)

7.4.2 Evaluation Results

We evaluate the effect of the closely coupled gaze-speech instances on word acquisition from the 1-best speech recognition. To show the influence of speech recognition quality on word acquisition performance, we also evaluate word acquisition from speech transcript. The predicted closely coupled gaze-speech instances in the evaluations are generated by a 10-fold cross validation with the logistic regression classifier.

Figures 7.2 ~ 7.7 show the precision, recall, and F-measure of the n-best words acquired by Model-2t-r using all instances (*all*), only predicted closely coupled instances (*predicted*), and true (manually labeled) closely coupled instances (*true*). In Figures 7.2 ~ 7.4, the acquired words come from the 1-best speech recognition of users' utterances. In Figures 7.5 ~ 7.7, the acquired words come from the transcripts of users' utterances.

Figure 7.8 compares the MRRRs achieved by Model-2t-r using different set of instances (all instances, predicted closely coupled instances, true closely coupled instances) with different speech input (1-best speech recognition, speech transcript).

Results of Word Acquisition on 1-best Speech Recognition

As shown in Figure 7.4, using predicted instances achieves consistent better performance than using all instances except the case where only the 3-best word candidates are evaluated. These results show that the prediction of closely coupled gaze-speech instances helps word acquisition. When the true closely coupled gaze-speech instances are used for word acquisition, the word acquisition performance is further improved. This means that higher gaze-speech prediction precision will lead to better word acquisition performance.

We notice that using all instances actually achieves higher F-measure than using

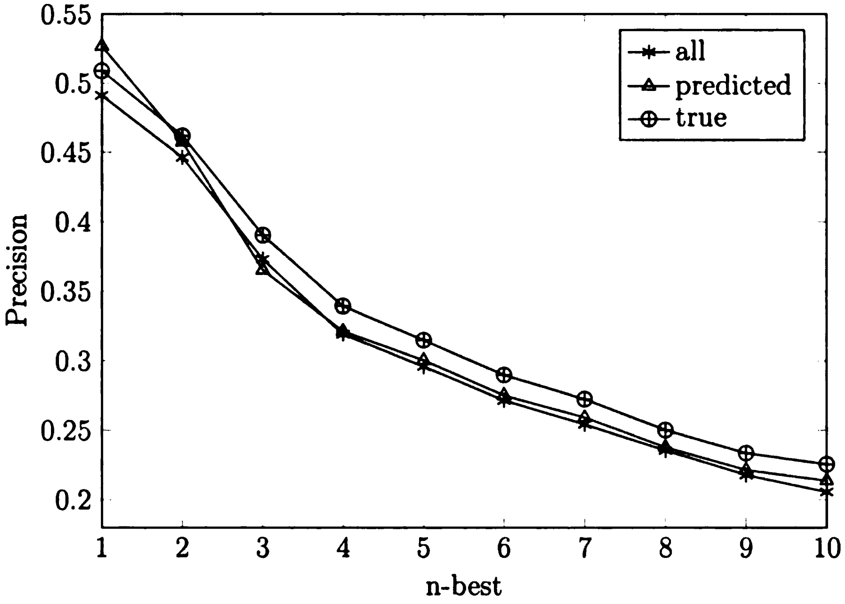


Figure 7.2. Precision of word acquisition on 1-best speech recognition with Model-2t-r

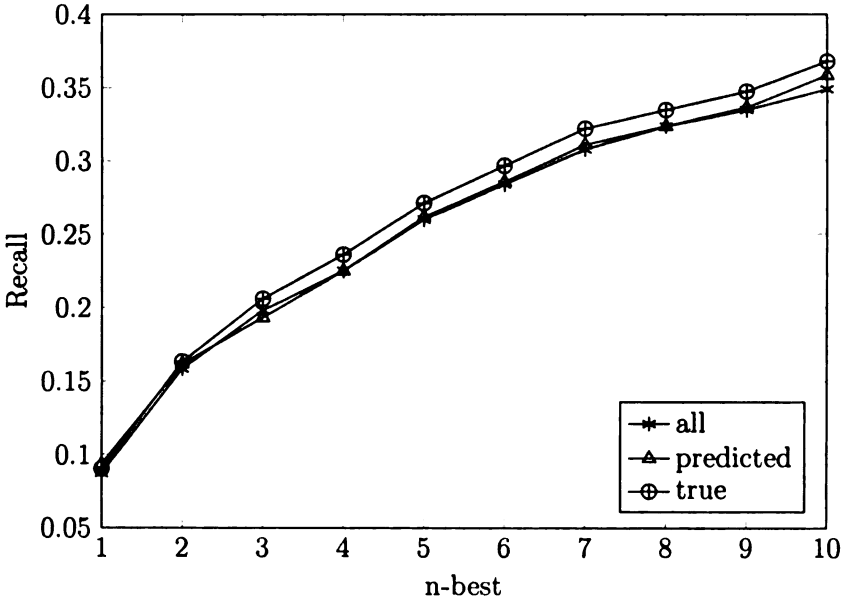


Figure 7.3. Recall of word acquisition on 1-best speech recognition with Model-2t-r

predicted instances for the 3-best word candidates. This is because there are few "gold standard" words that do not appear in the predicted gaze-speech instances due to the scarcity of these words in the whole data set. In the word acquisition with all

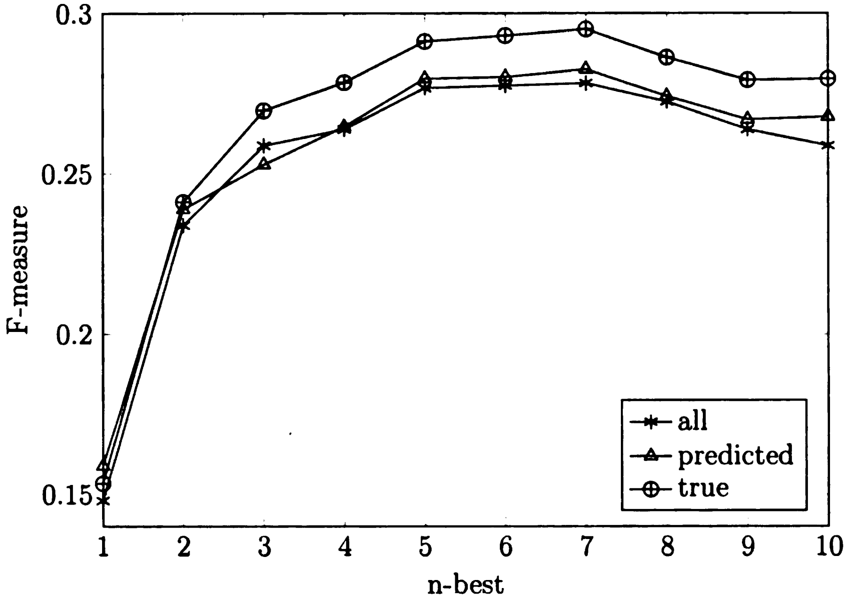


Figure 7.4. F-measure of word acquisition on 1-best speech recognition with Model-2t-r

instances, these words will not appear in the 10-best list if word acquisition is only based on co-occurring statistics (as in Model-1). In Model-2t-r, with domain semantic relatedness rescoring, these words are boosted up to the 3-best list. However, this can not happen in the word acquisition with the predicted gaze-speech instances because the predicted instances do not contain these few words and therefore it is impossible to acquire them. Therefore, for Model-2t-r, using all instances accidentally outperforms using predicted instances when only the 3-best word candidates are evaluated. We believe this will not happen when a fairly large amount of data is available for word acquisition.

As shown in Figure 7.8, the MRRRs achieved by Model-2t-r using different sets of instances with the 1-best speech recognition are consistent with their performances on F-measure. Using predicted instances results in significantly better MRRR than using all instances ($t = 1.89, p < 0.031$).

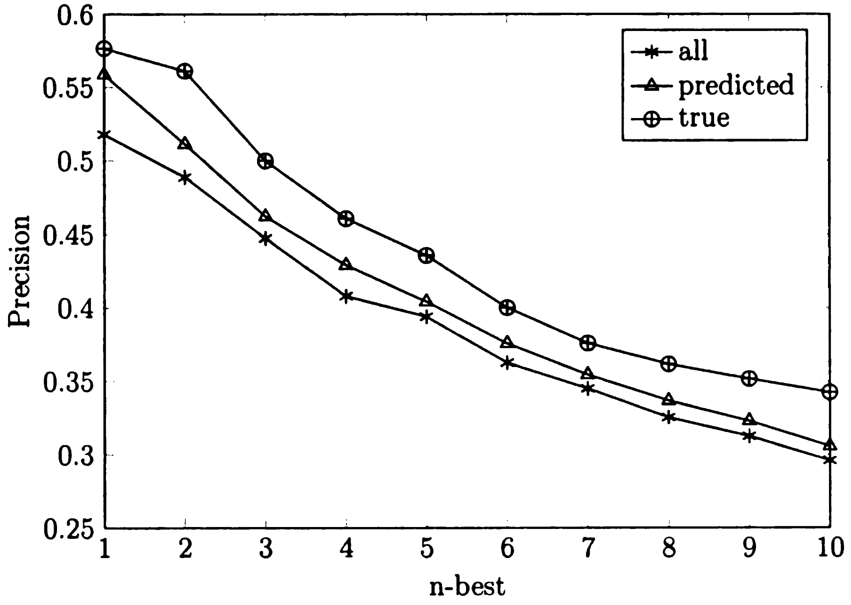


Figure 7.5. Precision of word acquisition on speech transcript with Model-2t-r

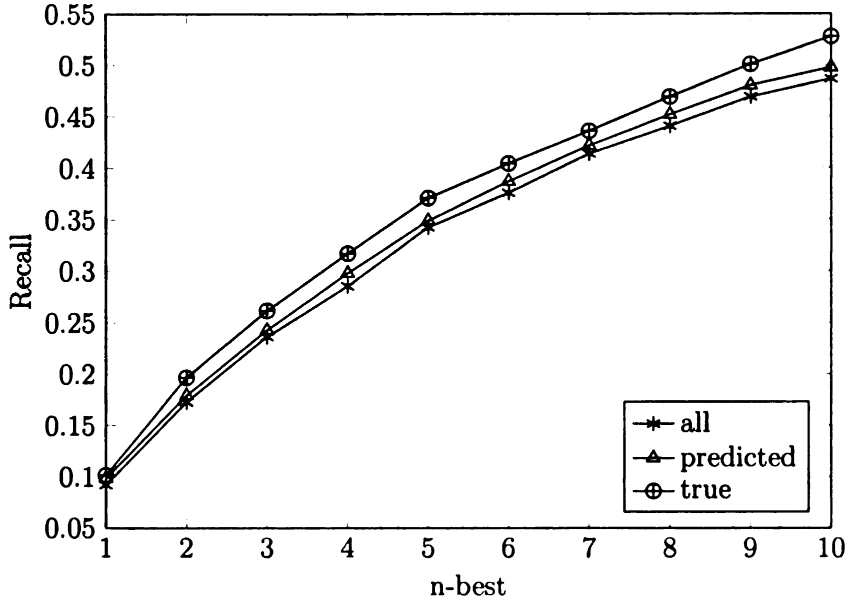


Figure 7.6. Recall of word acquisition on speech transcript with Model-2t-r

Results of Word Acquisition on Speech Transcript

For the word acquisition on speech transcript, as shown in Figure 7.7, using predicted closely coupled instances results in better F-measure than using all instances. When

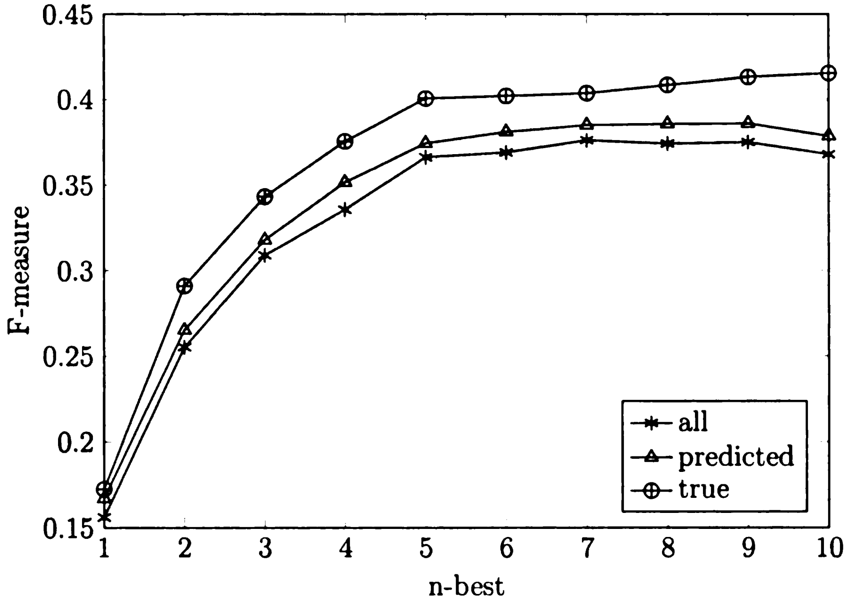


Figure 7.7. F-measure of word acquisition on speech transcript with Model-2t-r

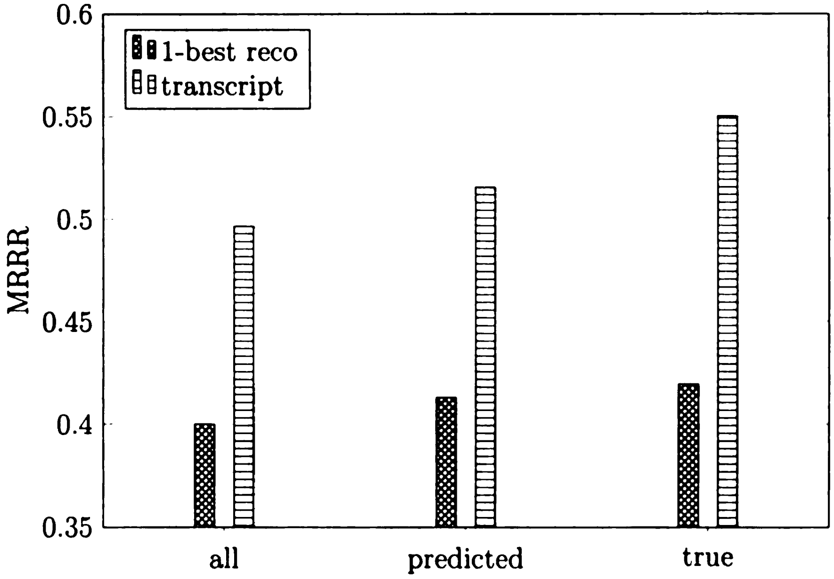


Figure 7.8. MRRRs achieved by Model-2t-r with different data set

the true closely coupled instances are used for word acquisition, the F-measure is further improved.

As shown in Figure 7.8, consistent with its F-measure performance, using predicted instances results in significantly better MRRR than using all instances (t =

$2.66, p < 0.005$).

The quality of speech recognition is critical to word acquisition performance. Figure 7.8 also compares the word acquisition performance on the 1-best speech recognition and speech transcript. As expected, the word acquisition performance on speech transcript is much better than on 1-best speech recognition. This result shows that better speech recognition will lead to better word acquisition.

7.5 The Effect of Word Acquisition on Language Understanding

One important goal of word acquisition is to use the acquired new words to help language understanding in subsequent conversation. To demonstrate the effect of online word acquisition on language understanding, we conduct simulation studies based on our collected data. In these simulations, the system starts with an initial knowledge base – a vocabulary of words associated to domain concepts. The system continuously enhances its knowledge base by acquiring words from users with Model-2t-r (Section 6.4) that incorporates both speech-gaze temporal information and domain semantic relatedness. The enhanced knowledge base is used to understand the language of new users.

We evaluate language understanding performance on concept identification rate (CIR):

$$CIR = \frac{\#\text{correctly identified conepts in the 1-best speech recognition}}{\#\text{concepts in the speech transcript}}$$

We simulate the process of online word acquisition and evaluate its effect on language understanding for two situations: 1) the system starts with no training data but with a small initial vocabulary, and 2) the system starts with some training data.

7.5.1 Simulation 1: When the System Starts with No Training Data

To build conversational systems, one approach is that domain experts provide domain vocabulary to the system at design time. Our first simulation follows this practice. The system is provided with a default vocabulary to start without training data. The default vocabulary contains one "seed" word for each domain concept.

Using the collected data of 20 users, the simulation process goes through the following steps:

- For user index $i = 1, 2, \ldots, 20$:
  - Evaluate CIR of the i-th user's utterances (1-best speech recognition) with the current system vocabulary.
  - Acquire words from all the instances (with 1-best speech recognition) of users $1 \cdots i$.
  - Among the 10-best acquired words, add verified new words to the system vocabulary.

In the above process, the language understanding performance on each individual user depends on the user's own language as well as the user's position in the user sequence. To reduce the effect of user ordering on language understanding performance, the above simulation process is repeated 500 times with randomly ordered users. The average of the CIRs in these simulations is shown in Figure 7.9.

Figure 7.9 also shows the CIRs when the system is with a static knowledge base (vocabulary). The curve is drawn in the same way as the curve with a dynamic knowledge base, except without word acquisition in the random simulation processes. As we can see in the figure, when the system doest not have word acquisition capability, its language understanding performance does not change after more users have communicated to the system. With the capability of automatic word acquisition, the

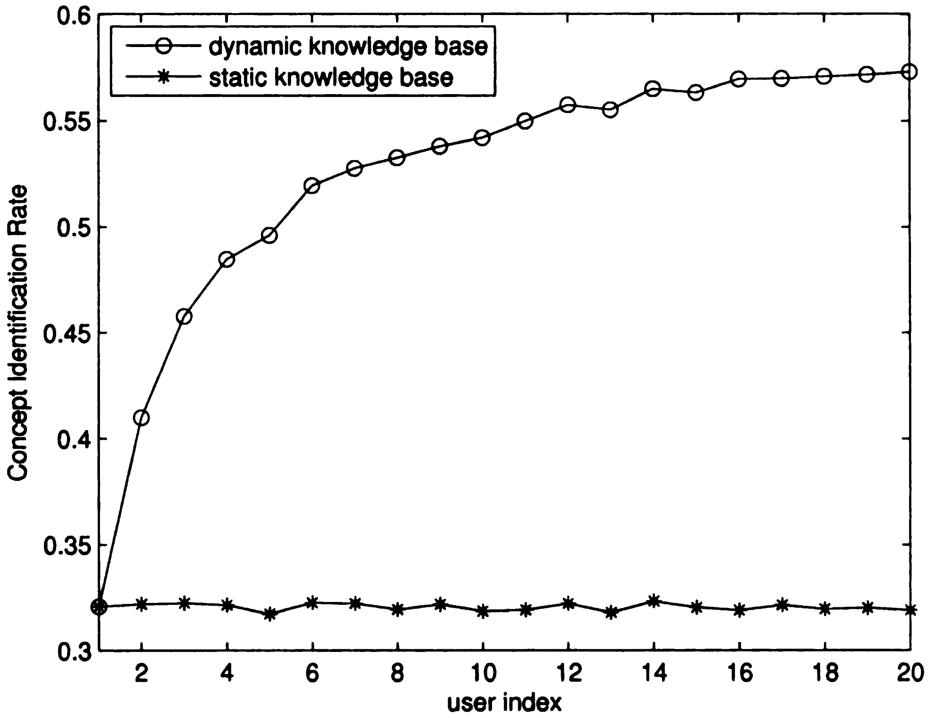


Figure 7.9. CIR of user language achieved by the system starting with no training data

system's language understanding performance becomes better after more users have talked to the system.

7.5.2 Simulation 2: When the System Starts with Training Data

Many conversational systems use real user data to derive domain vocabulary. To follow this practice, the second simulation provides the system with some training data. The training data serves two purposes: 1) build an initial vocabulary of the system; 2) train a classifier to predict the closely coupled gaze-speech instances of new users' data.

Using the collected data of 20 users, the simulation process goes through the following steps:

- Using the first m users' data as training data, acquire words from the training instances (with speech transcript); add the verified 10-best words to the system's vocabulary as "seed" words; build a classifier with the training data for prediction of closely coupled gaze-speech instances.

- Evaluate the effect of incremental word acquisition on CIR of the remaining (20-m) users' data. For user index $i = 1, 2, \ldots, (20\text{-}m)$:
  - Evaluate CIR of the i-th user's utterances (1-best speech recognition).
  - Predict coupled gaze-speech instances of the i-th user's data.
  - Acquire words from the m training users' true coupled instances (with speech transcript) and the predicted coupled instances (with 1-best speech recognition) of users $1 \cdots i$.
  - Among the 10-best acquired words, add verified new words to the system vocabulary.

The above simulation process is repeated 500 times with randomly ordered users to reduce the effect of user ordering on the language understanding performance. Figure 7.10 shows the averaged language understanding performance of these random simulations.

The language understanding performance of the system with a static knowledge base is also shown in Figure 7.10. The curve is drawn by the same random simulations without the steps of word acquisition. We can observe a general trend in the figure that, with word acquisition, the system's language understanding becomes better after more users have communicated to the system. Without word acquisition capability, the system's language understanding performance does not increase after more users have conversed with the system.

The simulations show that automatic vocabulary acquisition is beneficial to the system's language understanding performance when training data is available. When training data is not available, vocabulary acquisition could be more important and beneficial to robust language understanding. It is worth to mention that the results

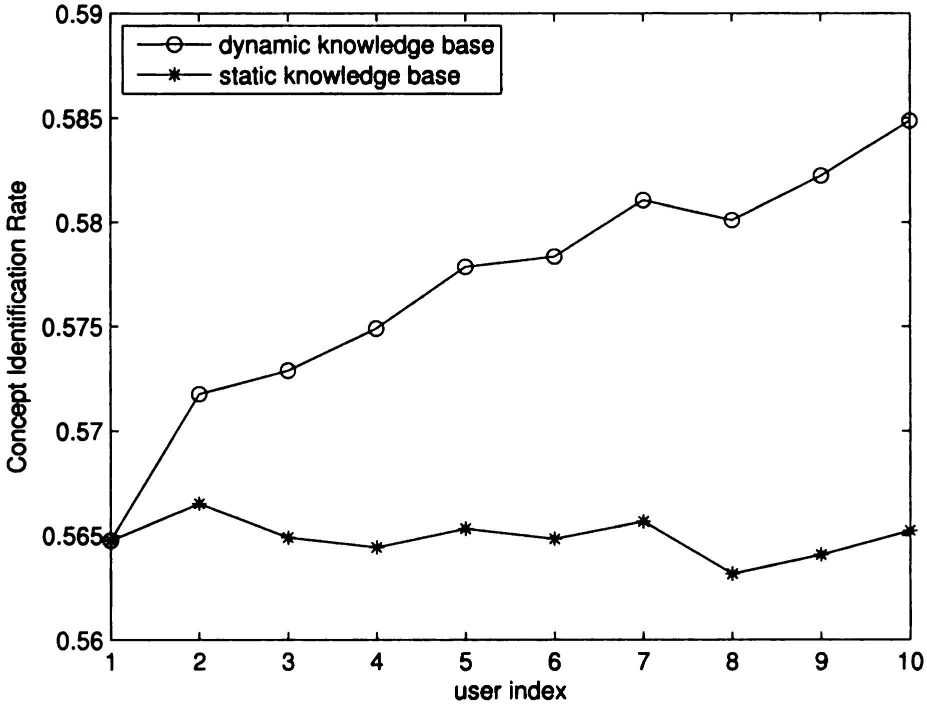


Figure 7.10. CIR of user language achieved by the system starting with 10 users training data)

shown here are based on the 1-best recognized speech hypotheses with a relatively high WER (48.1%). With better speech recognition, we expect to have better concept identification results.

7.6 Summary

This chapter investigates the automatic identification of closely coupled gaze-speech instances and its application for automatic word acquisition in multimodal conversational systems. Particulary, this chapter explores the use of the features extracted from speech, eye gaze, user interaction activities, and conversation context for predicting whether the user's naturally occurring eye gaze links to the content of the user's speech.

This chapter also investigates the application of the identified closely coupled gaze-speech instances for word acquisition The gaze-speech prediction and its effect on word acquisition are evaluated on the 1-best speech recognition and speech transcript. The experiments demonstrate that the automatic identification of the closely coupled gaze-speech instances significantly improves word acquisition, no matter the words are acquired from the 1-best speech recognition or from the speech transcript.

Moreover, this chapter demonstrates that, during multimodal conversation process, the system with word acquisition capability will be able to better understand the user's language after more users have communicated to the system.

CHAPTER 8

Conclusions

8.1 Contributions

In this thesis, we present our work on using non-verbal modalities for human language interpretation in multimodal conversational systems. Particularly, we present a joint solution to the problems of unreliable speech input and unexpected speech input in multimodal conversational systems, which includes two aspects: 1) use deictic gesture and eye gaze to improve speech recognition and understanding, and 2) use eye gaze to acquire new words automatically during multimodal conversation. Our evaluations have demonstrated the promise of incorporating non-verbal modalities to help speech recognition and language understanding during multimodal conversation.

Specific contributions of this thesis include:

- Systematic investigation of incorporating deictic gesture and eye gaze to improve speech recognition hypotheses for spoken language understanding. We have developed salience driven approaches to incorporate the domain context activated by gesture/gaze in speech recognition. The gesture/gaze-based salience driven language models are used in different stages of speech recognition to improve recognition hypotheses. Experimental results show that, by using non-verbal salience driven language models, the word error rate of speech recognition is decreased by 6.7% and the concept identification F-measure is increased by

4.2%.

- Systematic investigation of using deictic gesture to improve spoken language understanding in multimodal interpretation. We have developed model-based and instance-based approaches to incorporate gestural information in language understanding. Experimental results have shown that the accuracy of intention recognition in language understanding is increased by 6% $\sim$ 6.6% by different approaches that incorporate gestural information. We further analyze the implications of these results in building practical conversational systems.

- Systematic investigation of using eye gaze for automatic word acquisition in multimodal conversation. We have developed word acquisition models that incorporate speech-gaze temporal information and domain semantic relatedness to improve word acquisition. By using the temporal and semantic information, the mean reciprocal rank rate (MRRR) of word acquisition is increased by 43.2% in our experiment. To further improve word acquisition performance, we build a classifier based on user interactivity to pick out "useful" speech-gaze instances before word acquisition, which results in a further increase of MRRR by 3.6%. Our simulation studies have shown that automatic online word acquisition improves the system's language understanding performance.

- A Multimodal conversational system supporting speech, deictic gesture, and eye gaze developed for 3D domains. Integrating techniques from speech recognition, eye tracking, and computer graphics, we have implemented a multimodal conversational system based on 3D interior domains. The system can support speech, deictic gesture, and eye gaze inputs from the user during multimodal conversation. It provides a framework to develop different multimodal applications.

- Corpora of multimodal data collected through user studies. This research results in 3 sets of data to study multimodal conversation. These data provide user speech and the accompanying deictic gestures and eye gaze fixations during multimodal conversation. The data has been annotated for this thesis research. The annotation includes the transcript of speech, the timestamps of transcribed words, the referred entity in users' speech, and the labeling of closely-coupled gaze-speech pairs. These data will be available for research communities.

8.2 Future Directions

Some future directions for the research on using non-verbal modalities in language processing include:

- In this thesis's work on automatic word acquisition, new words are grounded to the domain concepts representing entities and their properties. These domain concepts are already given to the system. It is interesting for future work to automatically learn these domain concepts.

- The current implementation of word acquisition by means of eye gaze learns words referring to entities and their physical properties (color, size, materia, shape). It may be extended to learn words that describe the spatial relations of entities and the user actions.

- Besides word acquisition, eye gaze can also be used to help syntactic parsing of the user's spoken language. For example, suppose the user says "there is a book on a table with a brown cover". It is ambiguous in the parsing whether the prepositional phrase "with a brown cover" should be attached to "a book" or "a table". However, using eye gaze fixations, the system can decide which entity the phrase "with a brown cover" should be attached to based on its domain knowledge about the properties of the fixated entities (book, table).

APPENDICES

A Multimodal Data Collection

This section describes the user studies that we conducted to collect the speech-gesture and speech-gaze multimodal data sets for the investigations in this thesis.

A.1 Speech-Gesture Data Collection in the Interior Decoration Domain

We collected speech-gesture data by conducting user studies in the interior decoration domain (Section 3.3.1). In this study, users were asked to accomplish tasks in two scenarios. Scenario 1 was to clean up and redecorate a messy room. Scenario 2 was to arrange and decorate the room so that it looks like the room in the pictures provided to the user. Each scenario put the user into a specific role (e.g., college student, professor, merchant, etc.), and the task had to be completed with a set of constraints (e.g., budget of furnishings, bed size, number of domestic products, etc.). Figures A.1 & A.2 show the instructions for scenario 1 and scenario 2 that were given to the user before the study.

We recruited 5 users for the study. During the study, the user's speech was recorded through an open microphone and the user's deictic gesture was captured by a touch-screen. From the user studies, we collected 649 spoken utterances with accompanying gestures

A.2 Speech-Gaze Data Collection in the Interior Decoration Domain

We also collected a corpus of speech-gaze data in the interior decoration domain with a different user task. In this study, a static 3D bedroom scene was shown to the user. The system verbally asked the user a list of questions one at a time about the bedroom and the user answered the questions by speaking to the system. Figure A.3 lists the questions that are asked by the system.

We recruited 7 users for the study. During the study, the user's speech was recorded through an open microphone and the user's eye gaze was captured by an

Description of Scenario 1

1. You are planning to have an important meeting at your apartment. Currently, your apartment is a mess. You would like to clean it up and redecorate. You have found a computer program that will allow you to manipulate the furniture arrangement and style in the apartment. This will allow you to decorate the virtual replica of your apartment prior to redecorating your real apartment. This will minimize heavy lifting and save you lots of time! You have two goals. The first goal is to clean up your messy apartment by removing, replacing, or modifying objects that appear to be either out of place or have strange-looking characteristics. The second goal is to redecorate your apartment. This can be accomplished by adding, removing, or modifying objects.

2. You are not a millionaire, so you will have to stay under a specific budget during the decoration process. You also have certain personality traits and practical needs which will constrain the redecoration process. The budget along with these needs will be defined by a character role card which will be given to you at the beginning of this scenario.

3. Additionally you will need to write down certain information about the resulting redecorated apartment for future reference. The information that is important to you will be determined by your character role.

Role: College Student

You are a college student. You want to have an exotic and colorful apartment, but price is a major concern. You require a quality desk that will last for a long time. You need cabinets with many drawers to store all your school work. You prefer dim lighting and lots of plants and artwork. You want your apartment to look as exotic and colorful as possible while satisfying your basic needs and staying under a budget of $1800.

Role: Patriotic Family (with kids)

You are a former US Marine. You are very patriotic and have a family (with kids) that share your values. You want your apartment to contain as many objects made in the US (especially objects that have recently been made in the US) and be symbolic of the US, yet you also want your apartment to be practical and safe for your children. You prefer soft unbreakable furniture without sharp corners that has be recently been produced in the US. You need a large bed and would prefer to have at least one reclining piece of furniture. You must satisfy these preferences while staying under a budget of $2500.

Figure A.1. Instruction for scenario 1 in the interior decoration domain

Description of Scenario 2

1. Imagine that you are searching for a new place to live. You have found a computer program that will allow you to manipulate the furniture arrangement and style in a perspective apartment. When you recently visited an old friend, you really enjoyed the layout of his/her apartment. The images of this apartment are vividly engrained in your mind. Your goal is to arrange your perspective apartment in the mold of those images. To help with the story, sample images will be provided for you.

2. While the layout of your friend's place was aesthetically pleasing to you, certain aspects of the apartment need to be modified to fulfill your own personality traits and practical needs. These needs will be defined by a character role card which will be given to you at the beginning of this scenario. Based on your chosen character role, you will need to modify certain pieces of furniture to adhere to your character's needs.

3. Additionally you will need to write down certain information about the perspective apartment for future reference. The information that is important to you will be determined by your character role.

Role: Collector You are an art and antiques collector. You prefer old, expensive, and aesthetically pleasing furniture. You sometimes take prospective customers to your apartment and need to keep up the appearance that you know what you are talking about. Your goal for this apartment is that it contains a lot of art (paintings), old and expensive furniture, objects from a wide variety of countries with a minimal number of US-produced objects. You will need to modify the existing furniture to adhere to your preferences.

Role: Professor You are a college professor. The apartment's practicality is very important to you. You require a quality desk that will last for a long time. You need cabinets with many drawers. Light is very important to you. You prefer powerful (high-wattage) lamps. Additionally you require that a recliner is available when you need to relax from your busy day. You want to efficiently balance comfort vs. price – you generally don't want furniture made out of the cheapest or more expensive material.

Figure A.2. Instruction for scenario 2 in the interior decoration domain

Eye Link II eye tracker. From the user studies, we collected 554 spoken utterances with accompanying gaze streams.

1 Describe this room.
2 What do you like/dislike about the arrangement?
3 Describe anything in the room that seems strange to you.
4 Is there a bed in this room?
5 How big is the bed?
6 Describe the area around the bed.
7 Would you make any changes to the area around the bed?
8 Describe the left wall.
9 How many paintings are there in this room?
10 Which is your favorite painting?
11 Which is your least favorite painting?
12 What is your favorite piece of furniture in the room?
13 What is your least favorite piece of furniture in the room?
14 How would you change this piece of furniture to make it better?

Figure A.3. Questions for users in the study

A.3 Speech-Gaze Data Collection in the Treasure Hunting Domain

We collected another corpus of speech-gaze data by conducting user studies in the treasure hunting domain (Section 3.3.2). In this study, the user's task is to find some treasures that are hidden in a 3D castle. The user can walk around inside the castle and move objects. The user needs to consult with a remote "expert" (i.e., an artificial agent) to find the treasures. The expert has some knowledge about the treasures but can not see the castle. The user has to talk to the expert for advices of finding the treasures. Figure A.4 shows the instruction that is given to the user before the study.

We recruited 20 users for the study. During the study, the user's speech was recorded through an open microphone and the user's eye gaze was captured by a Tobii eye tracker. From the user studies, we collected 3709 spoken utterances with accompanying gaze streams.

Instruction

Your mission, if you choose to accept it (by signing the consent form), is to immerse yourself into the world of treasure hunting and find Zahalin's treasure. With the help of an artificial conversational agent, you will navigate Zahalin's castle in search for the treasure. Some of the treasure will be hidden, while some of it will be in plain sight. To communicate with your artificial assistant, speak clearly into the microphone using your natural tone of voice.

The assistant is an old criminal who is familiar with Zahalin's castle. He has partial knowledge about where the treasure is and how to find it, but cannot see what is inside the castle. You have additional knowledge about what can be seen in the castle environment. It is your responsibility to communicate with the artificial assistant and provide as much detail about the layout of the castle as the he requires.

You have the ability to open, move, and pick up various objects in the castle. However, you must be careful! Some objects are booby trapped and you will be penalized for manipulating these objects. Make sure to ask the artificial assistant if an object is safe before manipulating it.

Together you will decipher this puzzle. Good luck!

While you navigate through the castle and converse with your artificial assistant, we will track your speech and eye gaze. This data will be used to make further improvements to the conversational agent's spoken language understanding. The system will inform you if it fails to recognize either your speech or eye gaze. If this happens at any point during the study, please ask your proctor for assistance.

Figure A.4. Instruction for the user study

B Parameter Estimation in Approaches to Word Acquisition

Given parallel data set $(\mathbf{W}, \mathbf{E})$ where $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$ and $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, EM algorithms are used to estimate the probabilities $p(w|e)$ that maximize the likelihood of the data set

$$p(\mathbf{W}|\mathbf{E}) = \prod_{k=1}^{n} p(\mathbf{w}_k|\mathbf{e}_k)$$

B.1 Parameter Estimation for Base Model-1

The Base Model-1 is

$$p(\mathbf{w}|\mathbf{e}) = \frac{1}{(|\mathbf{e}|+1)^{|\mathbf{w}|}} \prod_{j=1}^{|\mathbf{w}|} \sum_{i=0}^{|\mathbf{e}|} p(w_j|e_i)$$

Use EM algorithm to estimate the parameters $\theta = \big(p(w|e)\big)$ that maximize $p(\mathbf{W}|\mathbf{E})$:

- E-step: compute the expected value of the log-likelihood with respect to the distribution of the alignments a_j

$$\begin{aligned} Q &= E\Big[\log p(\mathbf{W}|\mathbf{E}, \theta^{(old)})\Big] \\ &= \sum_{k=1}^{n} \log \frac{1}{(|\mathbf{e}_k|+1)^{|\mathbf{w}_k|}} \\ &\quad + \sum_{k=1}^{n} \sum_{j=1}^{|\mathbf{w}_k|} \sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)}) \log p(w_{kj}|e_{ki}) \end{aligned}$$

  where for each instance,

$$p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)}) = \frac{p(w_{kj}|e_{ki})}{\sum_{i=0}^{|\mathbf{e}_k|} p(w_{kj}|e_{ki})} \qquad \text{(B.1)}$$

- M-step: find the new parameters

$$\theta^{(new)} = \arg\max_{\theta} Q$$

and we have

$$p(w|e) = \frac{\sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})\delta(w, w_{kj})\delta(e, e_{ki})}{\sum_{w}\sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})\delta(w, w_{kj})\delta(e, e_{ki})} \qquad \text{(B.2)}$$

where,

$$\delta(w, w_{kj}) = \begin{cases} 1 & w_{kj} = w \\ 0 & \text{otherwise} \end{cases} \qquad \text{(B.3)}$$

$$\delta(e, e_{ki}) = \begin{cases} 1 & e_{ki} = e \\ 0 & \text{otherwise} \end{cases} \qquad \text{(B.4)}$$

B.2 Parameter Estimation for Base Model-2

The Base Model-2 is

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{|\mathbf{w}|}\sum_{i=0}^{|\mathbf{e}|} p(a_j = i|j, |\mathbf{w}|, |\mathbf{e}|)p(w_j|e_i)$$

Use EM algorithm to estimate the parameters $\theta = \big(p(a_j|m, l), p(w|e)\big)$ that maximize $p(\mathbf{W}|\mathbf{E})$:

- E-step: compute the expected value of the log-likelihood with respect to the distribution of the alignments a_j

$$\begin{aligned} Q &= E\Big[\log p(\mathbf{W}|\mathbf{E}, \theta^{(old)})\Big] \\ &= \sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, |\mathbf{w}_k|, |\mathbf{e}_k|, \theta^{(old)}) \\ &\qquad\qquad \times \log\big[p(a_j = i||\mathbf{w}_k|, |\mathbf{e}_k|)p(w_{kj}|e_{ki})\big] \end{aligned}$$

where for each instance k,

$$p(a_j = i|w_{kj}, e_{ki}, |\mathbf{w}_k|, |\mathbf{e}_k|, \theta^{(old)}) = \frac{p(a_j = i||\mathbf{w}_k|, |\mathbf{e}_k|)p(w_{kj}|e_{ki})}{\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i||\mathbf{w}_k|, |\mathbf{e}_k|)p(w_{kj}|e_{ki})} \qquad \text{(B.5)}$$

- M-step: find the new parameters

$$\theta^{(new)} = \arg\max_{\theta} Q$$

and we have

$$p(a_j = i|m,l) = \frac{\displaystyle\sum_{k:|\mathbf{w}_k|=m,|\mathbf{e}_k|=l} p(a_j = i|w_{kj}, e_{ki}, |\mathbf{w}_k|, |\mathbf{e}_k|, \theta^{(old)})}{\displaystyle\sum_{k:|\mathbf{w}_k|=m,|\mathbf{e}_k|=l} 1} \tag{B.6}$$

$$p(w|e) = \frac{\displaystyle\sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|m = |\mathbf{w}_k|, l = |\mathbf{e}_k|)\delta(w, w_{kj})\delta(e, e_{ki})}{\displaystyle\sum_{w}\sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|m = |\mathbf{w}_k|, l = |\mathbf{e}_k|)\delta(w, w_{kj})\delta(e, e_{ki})} \tag{B.7}$$

where $\delta(w, w_{kj})$ and $\delta(e, e_{ki})$ are shown in Equations B.3 and B.4.

B.3 Parameter Estimation for Model-2s

The Model-2s is

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{|\mathbf{w}|}\sum_{i=0}^{|\mathbf{e}|} p_s(a_j = i|j, \mathbf{e}, \mathbf{w})p(w_j|e_i)$$

Use EM algorithm to estimate the parameters $\theta = \big(p(w|e)\big)$ that maximize $p(\mathbf{W}|\mathbf{E})$:

- E-step: compute the expected value of the log-likelihood with respect to the distribution of the alignments a_j

$$\begin{aligned} Q &= E\Big[\log p(\mathbf{W}|\mathbf{E}, \theta^{(old)})\Big] \\ &= \sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)}) \log\big[p_s(a_j = i|j, \mathbf{e}_k, \mathbf{w}_k)p(w_{kj}|e_{ki})\big] \end{aligned}$$

where for each instance,

$$p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)}) = \frac{p_s(a_j = i|j, \mathbf{e}_k, \mathbf{w}_k)p(w_{kj}|e_{ki})}{\displaystyle\sum_{i=0}^{|\mathbf{e}_k|} p_s(a_j = i|j, \mathbf{e}_k, \mathbf{w}_k)p(w_{kj}|e_{ki})} \tag{B.8}$$

- M-step: find the new parameters

$$\theta^{(new)} = \arg\max_{\theta} Q$$

and we have

$$p(w|e) = \frac{\sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})\delta(w, w_{kj})\delta(e, e_{ki})}{\sum_{w}\sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})\delta(w, w_{kj})\delta(e, e_{ki})} \tag{B.9}$$

where $\delta(w, w_{kj})$ and $\delta(e, e_{ki})$ are shown in Equations B.3 and B.4.

B.4 Parameter Estimation for Model-2t

The Model-2t is

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{|\mathbf{w}|}\sum_{i=0}^{|\mathbf{e}|} p_t(a_j = i|j, \mathbf{e}, \mathbf{w})p(w_j|e_i)$$

where

$$p_t(a_j = i|j, \mathbf{e}, \mathbf{w}) = \begin{cases} 0 & d(e_i, w_j) > 0 \\ \dfrac{\exp[\alpha \cdot d(e_i, w_j)]}{\sum_i \exp[\alpha \cdot d(e_i, w_j)]} & d(e_i, w_j) \leq 0 \end{cases}$$

Use EM algorithm to estimate the parameters $\theta = \big(p(w|e), \alpha\big)$ that maximize $p(\mathbf{W}|\mathbf{E})$:

- E-step: compute the expected value of the log-likelihood with respect to the distribution of the alignments a_j

$$\begin{aligned} Q &= E\Big[\log p(\mathbf{W}|\mathbf{E}, \theta^{(old)})\Big] \\ &= \sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)}) \log\big[p_t(a_j = i|j, \mathbf{e}_k, \mathbf{w}_k)p(w_{kj}|e_{ki})\big] \end{aligned}$$

where for each instance,

$$\begin{aligned} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)}) &= \frac{p_t(a_j = i|j, \mathbf{e}_k, \mathbf{w}_k)p(w_{kj}|e_{ki})}{\sum\limits_{i=0}^{|\mathbf{e}_k|} p_t(a_j = i|j, \mathbf{e}_k, \mathbf{w}_k)p(w_{kj}|e_{ki})} \\ &= \frac{\exp[\alpha \cdot d(e_{ki}, w_{kj})]p(w_{kj}|e_{ki})}{\sum\limits_{i=0}^{|\mathbf{e}_k|} \exp[\alpha \cdot d(e_{ki}, w_{kj})]p(w_{kj}|e_{ki})} \end{aligned} \tag{B.10}$$

- M-step: find the new parameters

$$\theta^{(new)} = \arg\max_{\theta} Q$$

The new $p(w|e)$ is given by

$$p(w|e) = \frac{\sum\limits_{k=1}^{n}\sum\limits_{j=1}^{|\mathbf{w}_k|}\sum\limits_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})\delta(w, w_{kj})\delta(e, e_{ki})}{\sum\limits_{w}\sum\limits_{k=1}^{n}\sum\limits_{j=1}^{|\mathbf{w}_k|}\sum\limits_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})\delta(w, w_{kj})\delta(e, e_{ki})} \tag{B.11}$$

where $\delta(w, w_{kj})$ and $\delta(e, e_{ki})$ are shown in Equations B.3 and B.4.

The new α is given by

$$\frac{\exp[\alpha \cdot d(e_{ki}, w_{kj})]}{\sum\limits_{i} \exp[\alpha \cdot d(e_{ki}, w_{kj})]} = p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})$$

$$\alpha = \arg\min_{\alpha} \sum_{i}\sum_{j}\sum_{k}\left[\frac{\exp[\alpha \cdot d(e_{ki}, w_{kj})]}{\sum\limits_{i} \exp[\alpha \cdot d(e_{ki}, w_{kj})]} - p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})\right]^2 \tag{B.12}$$

The Levenberg-Marquardt (LM) algorithm [63, 67] is used to find the MSE estimate of α.

B.5 Parameter Estimation for Model-2ts

The Model-2ts is

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{|\mathbf{w}|} \sum_{i=0}^{|\mathbf{e}|} p_{ts}(a_j = i|j, \mathbf{e}, \mathbf{w}) p(w_j|e_i)$$

where

$$p_{ts}(a_j = i|j, \mathbf{e}, \mathbf{w}) = \frac{SR(e_{ki}, w_{kj}) \exp[\alpha \cdot d(e_i, w_j)]}{\sum_i SR(e_{ki}, w_{kj}) \exp[\alpha \cdot d(e_i, w_j)]}$$

Use EM algorithm to estimate the parameters $\theta = \big(p(w|e), \alpha\big)$ that maximize $p(\mathbf{W}|\mathbf{E})$:

- E-step: compute the expected value of the log-likelihood with respect to the distribution of the alignments a_j

$$\begin{aligned} Q &= E\Big[\log p(\mathbf{W}|\mathbf{E}, \theta^{(old)})\Big] \\ &= \sum_{k=1}^{n} \sum_{j=1}^{|\mathbf{w}_k|} \sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)}) \log\big[p_{ts}(a_j = i|j, \mathbf{e}_k, \mathbf{w}_k) p(w_{kj}|e_{ki})\big] \end{aligned}$$

where for each instance,

$$\begin{aligned} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)}) &= \frac{p_{ts}(a_j = i|j, \mathbf{e}_k, \mathbf{w}_k) p(w_{kj}|e_{ki})}{\sum_{i=0}^{|\mathbf{e}_k|} p_{ts}(a_j = i|j, \mathbf{e}_k, \mathbf{w}_k) p(w_{kj}|e_{ki})} \\ &= \frac{SR(e_{ki}, w_{kj}) \exp[\alpha \cdot d(e_{ki}, w_{kj})] p(w_{kj}|e_{ki})}{\sum_{i=0}^{|\mathbf{e}_k|} SR(e_{ki}, w_{kj}) \exp[\alpha \cdot d(e_{ki}, w_{kj})] p(w_{kj}|e_{ki})} \end{aligned} \tag{B.13}$$

- M-step: find the new parameters

$$\theta^{(new)} = \arg\max_{\theta} Q$$

The new $p(w|e)$ is given by

$$p(w|e) = \frac{\sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})\delta(w, w_{kj})\delta(e, e_{ki})}{\sum_{w}\sum_{k=1}^{n}\sum_{j=1}^{|\mathbf{w}_k|}\sum_{i=0}^{|\mathbf{e}_k|} p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})\delta(w, w_{kj})\delta(e, e_{ki})} \tag{B.14}$$

where $\delta(w, w_{kj})$ and $\delta(e, e_{ki})$ are shown in Equations B.3 and B.4.

The new α is given by

$$\frac{SR(e_{ki}, w_{kj}) \exp[\alpha \cdot d(e_{ki}, w_{kj})]}{\sum_i SR(e_{ki}, w_{kj}) \exp[\alpha \cdot d(e_{ki}, w_{kj})]} = p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)})$$

$$\alpha = \arg\min_{\alpha} \sum_i \sum_j \sum_k \left[\frac{SR(e_{ki}, w_{kj}) \exp[\alpha \cdot d(e_{ki}, w_{kj})]}{\sum_i SR(e_{ki}, w_{kj}) \exp[\alpha \cdot d(e_{ki}, w_{kj})]} - p(a_j = i|w_{kj}, e_{ki}, \theta^{(old)}) \right]^2 \tag{B.15}$$

The Levenberg-Marquardt (LM) algorithm [63, 67] is used to find the MSE estimate of α.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] J. F. Allen, B. W. Miller, E. K. Ringger, and T. Sikorski. A robust system for natural spoken dialogue. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1996.

[2] P. D. Allopenna, J. S. Magnuson, and M. K. Tanenhaus. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory & Language*, 38:419–439, 1998.

[3] K. Bock, D. E. Irwin, D. J. Davidson, and W. Leveltb. Minding the clock. *Journal of Memory and Language*, 48:653–685, 2003.

[4] R. A. Bolt. Put that there: Voice and gesture at the graphics interface. *Computer Graphics*, 14(3):262–270, 1980.

[5] P. F. Brown, S. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[6] P. F. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[7] S. Brown-Schmidt and M. K. Tanenhaus. Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, 54:592–609, 2006.

[8] E. Campana, J. Baldridge, J. Dowding, B. Hockey, R. Remington, and L. Stone. Using eye movements to determine referents in a spoken dialogue system. In *Proceedings of the Workshop on Perceptive User Interface*, 2001.

[9] S. Carbini, J. E. Viallet, and L. Delphin-Poulat. Context dependent interpretation of multimodal speech-pointing gesture interface. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, 2005.

[10] J. Chai, P. Hong, M. Zhou, and Z. Prasov. Optimization in multimodal interpretation. In *Proceedings of 42nd Annual Meeting of Association for Computational Linguistics (ACL)*, 2004.

[11] J. Chai, S. Pan, and M. Zhou. MIND: A context-based multimodal interpretation framework in conversational systems. In O. Bernsen, L. Dybkjaer, and J. van Kuppevelt, editors, *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems.* Kluwer Academic Publishers, 2005.

[12] J. Chai, Z. Prasov, and S. Qu. Cognitive principles in robutst multimodal interpretation. *Journal of Artificial Intelligence Research*, 27:55–83, 2006.

[13] J. Chai and S. Qu. A salience driven approach to robust input interpretation in multimodal conversational systems. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.

[14] J. Y. Chai, P. Hong, and M. X. Zhou. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, pages 70–77, 2004.

[15] A. Cheyer and L. Julia. MVIEWS: Multimodal tools for the video analyst. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, 1998.

[16] A. Chotimongkol and A. Rudnicky. N-best speech hypotheses reordering using linear regression. In *Proceedings of 7th EUROSPEECH*, pages 1829–1832, 2001.

[17] J. Chu-Carroll. MIMIC: An adaptive mixed initiative spoken dialogue system for information queries. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP)*, 2000.

[18] CMU. The CMU audio databases. http://www.speech.cs.cmu.edu/databases/.

[19] M. Coen, L. Weisman, K. Thomas, and M. Groh. A context sensitive natural language modality for the intelligent room. In *Proceedings of the 1st International Workshop on Managing Interactions in Smart Environments (MANSE)*, pages 38–79, 1999.

[20] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. QuickSet: multimodal interaction for distributed applications. In *Proceedings of the Fifth ACM International Conference on Multimedia*, pages 31–40, 1997.

[21] N. J. Cooke. *Gaze-Contingent Automatic Speech Recognition.* PhD thesis, University of Birminham, 2006.

[22] C. Cortes and V. Vapnik. Support-vector network. *Machine Learning*, 20:273–297, 1995.

[23] D. Dahan and M. K. Tanenhaus. Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review*, 12(3):453–459, 2005.

[24] S. Dupont and J. Luettin. Audio-visual speech modelling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151, 2000.

[25] S. Dusan, G. J. Gadbois, and J. Flanagan. Multimodal interaction on pda's integrating speech and pen inputs. In *Proceeding of EUROSPEECH*, 2003.

[26] K. M. Eberhard, M. J. Spivey-Knowiton, J. C. Sedivy, and M. K. Tanenhaus. Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24:409–436, 1995.

[27] J. Eisenstein and C. M. Christoudias. A salience-based approach to gesture-speech alignment. In *Proceedings of HLT/NAACL'04*, 2004.

[28] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

[29] P. Gorniak and D. Roy. Probabilistic grounding of situated speech using plan recognition and reference resolution. In *Proceedings of the Seventh International Conference on Multimodal Interfaces (ICMI)*, 2005.

[30] Z. M. Griffin. Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, 82:B1–B14, 2001.

[31] Z. M. Griffin and K. Bock. What the eyes say about speaking. *Psychological Science*, 11:274–279, 2000.

[32] B. J. Grosz, A. K. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226, 1995.

[33] B. J. Grosz and C. Sidner. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[34] A. Gruenstein, C. Wang, and S. Seneff. Context-sensitive statistical language modeling. In *Proceedings of Eurospeech*, 2005.

[35] J. K. Gundel, N. Hedberg, and R. Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307, 1993.

[36] J. E. Hanna and M. K. Tanenhaus. Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science*, 28:105–115, 2004.

[37] J. M. Henderson and F. Ferreira, editors. *The interface of language, vision, and action: Eye movements and the visual world.* Taylor & Francis, New York, 2004.

[38] H. Holzapfel, K. Nickel, and R. Stiefelhagen. Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures. In *Proceedings of the 6th international conference on Multimodal interfaces (ICMI)*, pages 175–182, 2004.

[39] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.

[40] P. Hui and H. Meng. Joint interpretation of input speech and pen gestures for multimodal human computer interaction. In *Proceedings of Interspeech*, 2006.

[41] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 16(5):550–554, 1994.

[42] C. Huls, E. Bos, and W. Classen. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics*, 21(1):59–79, 1995.

[43] R. J. K. Jacob. The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems*, 9(3):152–169, 1991.

[44] M. Johnston. Unification-based multimodal parsing. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, 1998.

[45] M. Johnston and S. Bangalore. Finite-state multimodal parsing and understanding. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2000.

[46] M. Johnston and S. Bangalore. Finite-state methods for multimodal parsing and integration. In *ESSLLI Workshop on Finite-state Methods*, 2001.

[47] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. MATCH: An architecture for multimodal dialogue systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 376–383, 2002.

[48] M. Johnston, P. Cohen, D. McGee, S. Oviatt, J. Pittman, and I. Smith. Unification-based multimodal integration. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1997.

[49] M. A. Just and P. A. Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, 8:441–480, 1976.

[50] D. Kahneman. *Attention and Effort.* Prentice-Hall, Inc., Englewood Cliffs, 1973.

[51] S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 35(3):400–401, 1987.

[52] M. Kaur, M. Termaine, N. Huang, J. Wilder, Z. Gacovski, F. Flippo, and C. S. Mantravadi. Where is "it"? event synchronization in gaze-speech input systems. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, 2003.

[53] Z. Kazi, S. Chen, M. Beitler, D. Chester, and R. Foulds. Multimodal HCI for robot control: Towards an intelligent robotic assistant f or people with disablities. In *Proceedings of AAAI'96 Fall Symposium on Developing AI Applications for the Disabled*, 1996.

[54] A. Kehler. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 685–689, 2000.

[55] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL)*, 2003.

[56] D. B. Koons, C. J. Sparrell, and K. R. Thorisson. Integrating simultaneous input from speech, gaze, and hand gestures. In M. Maybury, editor, *Intelligent Multimedia Interfaces*, pages 257–276. MIT Press, 1993.

[57] F. Landragin, N. Bellalem, and L. Romary. Visual salience and perceptual grouping in multimodal interactivity. In *Proceedings of the First International Workshop on Information Presentation and Natural Multimodal Dialogue*, pages 151–155, 2001.

[58] S. Lappin and H. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

[59] M. E. Latoschik. A user interface framework for multimodal vr interactions. In *Proceedings of the 7th international conference on Multimodal interfaces (ICMI)*, pages 76–83, 2005.

[60] S. le Cessie and J. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[61] Y. LeCun and C. Cortes. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist.

[62] O. Lemon and A. Gruenstein. Multithreaded context for robust conversational interfaces: Context-sensitive speech recognition and interpretation of corrective fragments. *ACM Transactions on Computer-Human Interaction*, 11(3):241–267, 2004.

[63] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, 1944.

[64] E. Levin, S. Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. D. Fabbrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P.Ruscitti, and M. Walker. The at&t-darpa communicator mixed-initiative spoken dialog system. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2000.

[65] D. J. Litman and K. Forbes-Riley. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

[66] Y. Liu, J. Y. Chai, and R. Jin. Automated vocabulary acquisition and interpretation in multimodal conversational systems. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2007.

[67] D. Marquardt. An algorithm for the least-squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics*, 11(2):431C441, 1963.

[68] A. S. Meyer, A. M. Sleiderink, and W. J. M. Levelt. Viewing and naming objects: eye movements during noun phrase production. *Cognition*, 66(22):25–33, 1998.

[69] L.-P. Morency and T. Darrell. Head gesture recognition in intelligent interfaces: The role of context in improving recognition. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, 2006.

[70] Y. Nakano, G. Reinstein, T. Stocky, and J. Cassell. Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003.

[71] J. G. Neal and S. C. Shapiro. Intelligent multimedia interface technology. In J. Sullivan and S. Tyler, editors, *Intelligent User Interfaces.* ACM, New York, 1991.

[72] J. G. Neal, C. Y. Thielman, Z. H. Dobes, S. M., and S. C. Shapiro. Natural language with integrated deictic and graphic gestures. In M. Maybury and W. Wahlster, editors, *Intelligent User Interfaces*, pages 38–51. Morgan Kaufmann Press, CA, 1998.

[73] S. Oviatt. Mulitmodal interactive maps: Designing for human performance. *Human-Computer Interaction*, 12:93–129, 1997.

[74] S. Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 1999.

[75] S. Oviatt. Breaking the robustness barrier: Recent progress on the design of robust multimodal systems. *Advances in Computers*, 56:305–341, 2002.

[76] S. Oviatt. Multimodal interfaces. In J. Jacko and A. Sears, editors, *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, chapter 14, pages 286–304. Lawrence Erlbaum Assoc., Mahwah, NJ, 2003.

[77] T. Pedersen, S. Patwardhan, and J. Michelizzi. WordNet::Similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI)*, 2004.

[78] A. Potamianos, S. Narayanan, and G. Riccardi. Adapative categorical understanding for spoken dialog systems. *IEEE Transactions on Speech and Audio Processing*, 13(3):321–329, 2005.

[79] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. In G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, editors, *Issues in Visual and Audio-Visual Speech Processing.* MIT Press, 2004.

[80] Z. Prasov and J. Y. Chai. What's in a gaze? the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of ACM 12th International Conference on Intelligent User interfaces (IUI)*, 2008.

[81] S. Qu and J. Y. Chai. Salience modeling based on non-verbal modalities for spoken language understanding. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 193–200, 2006.

[82] S. Qu and J. Y. Chai. An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 284–291, 2007.

[83] S. Qu and J. Y. Chai. Beyond attention: The role of deictic gesture in intention recognition in multimodal conversational interfaces. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, pages 237–246, 2008.

[84] S. Qu and J. Y. Chai. Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 244–253, 2008.

[85] S. Qu and J. Y. Chai. Speech-gaze temporal alignment for automatic word acquisition in multimodal conversational systems. In *Proceedings of the Fifth Midwest Computational Linguistics Colloquium (MCLC)*, 2008.

[86] R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[87] P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2005.

[88] K. Rayner. Eye movements in reading and information processing - 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.

[89] D. Roy and N. Mukherjee. Towards situated speech understanding: Visual context priming of language models. *Computer Speech and Language*, 19(2):227–248, 2005.

[90] D. K. Roy and A. P. Pentland. Learning words from sights and sounds, a computational model. *Cognitive Science*, 26(1):113–146, 2002.

[91] A. Sankar and A. Gorin. Adaptive language acquisition in a multi-sensory device. In R. Mammone, editor, *Artificial neural networks for speech and vision*, pages 324–356. Chapman and Hall, London, 1993.

[92] S. Seneff, D. Goddeau, C. Pao, and J. Polifroni. Multimodal discourse modelling in a multi-user multi-domain environment. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 192–195, 1996.

[93] A. Shaikh, S. Juth, A. Medl, I. Marsic, C. Kulikowski, and J. Flanagan. An architecture for multimodal information fusion. In *Proceedings of the Workshop on Perceptual User Interfaces (PUI)*, pages 91–93, 1997.

[94] P. Silsbee and A. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351, 1996.

[95] J. Siroux, M. Guyomard, F. Multon, and C. Remondeau. Modeling and processing of oral and tactile activities in the GEORAL system. In *Multimodal Human-Computer Communication, Systems, Techniques, and Experiments*, pages 101–110. Springer-Verlag, London, UK, 1998.

[96] R. A. Solsona, E. Fosler-Lussier, H.-K. J. Kuo, A. Potamianos, and I. Zitouni. Adaptive language models for spoken dialogue systems. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[97] M. J. Spivey, M. K. Tanenhaus, K. M. Eberhard, and J. C. Sedivy. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45:447–481, 2002.

[98] R. Stevenson. The role of salience in the production of referring expressions: A psycholinguistic perspective. In K. van Deemter and R. Kibble, editors, *Information Sharing*. CSLI Publ., 2002.

[99] Y. Sun, F. Chen, Y. Shi, and V. Chung. An input-parsing algorithm supporting integration of deictic gesture in natural language interface. In J. A. Jacko, editor, *Human-Computer Interaction: HCI Intelligent Multimodal Interaction Environments*, pages 206–215. Springer-Verlag Berlin Heidelberg, 2007.

[100] K. Tanaka. A robust selection system using real-time multi-modal user-agent interactions. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, 1999.

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

[102] M. J. Tomlinson, M. J. Russell, and N. M. Brooke. Integrating audio and visual information to provide highly robustspeech recognition. In *International*

Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 821–824, 1996.

[103] B. M. Velichkovsky. Communicating attention-gaze position transfer in cooperative problem solving. *Pragmatics and Cognition*, 3:99–224, 1995.

[104] J. Vergo. A statistical approach to multimodal natural language interaction. In *Proceedings of the AAAI'98 Workshop on Representations for Multimodal Human-Computer Interaction*, pages 81–85, 1998.

[105] R. Vertegaal. The GAZE groupware system: Mediating joint attention in multiparty communication and collaboration. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 294–301, 1999.

[106] M. T. Vo and C. Wood. Building an application framework for speech and pen input integration in multimodal learning interfaces. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996.

[107] W. Wahlster. User and discourse models for multimodal communication. In J. W. Sullivan and S. W. Tyler, editors, *Intelligent user interfaces*, pages 45–67. ACM, 1991.

[108] A. Waibel, B. Suhm, M. Vo, and J. Yang. Multimodal interfaces for multimedia information agents. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 167–170, 1997.

[109] A. Waibel, M. T. Vo, P. Duchnowski, and S. Manke. Multimodal interfaces. *Artificial Intelligence Review*, 10(3-4):299–319, 1996.

[110] M. A. Walker. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416, 2000.

[111] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel. Sphinx-4: A flexible open source framework for speech recognition. Technical Report TR-2004-139, Sun Microsystems Laboratories, 2004.

[112] J. Wang. Integration of eye-gaze, voice and manual response in multimodal user interfaces. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, pages 3938–3942, 1995.

[113] C. Ware and H. H. Mikaelian. An evaluation of an eye tracker as a device for computer input2. In *Proceedings of the SIGCHI/GI conference on Human factors in computing systems and graphics interface*, pages 183–188, 1987.

[114] Y. Watanabe, K. Iwata, R. Nakagawa, K. Shinoda, and S. Furui. Semi-synchronous speech and pen input. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[115] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2005.

[116] L. Wu, S. Oviatt, and P. Cohen. From members to teams to committee - a robust approach to gestural and multimodal recognition. *IEEE Transactions on Neural Networks*, 13(4), 2002.

[117] C. Yu and D. H. Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perceptions*, 1(1):57–80, 2004.

[118] C. Yu and D. H. Ballard. On the integration of grounding language and learning objects. In *Proceedings of AAAI-04*, 2004.

[119] M. Zancanaro, O. Stock, and C. Strapparava. Multimodal interaction for information access: Exploiting cohesion. *Computational Intelligence*, 13(7):439–464, 1997.

[120] S. Zhai, C. Morimoto, and S. Ihde. Manual and gaze input cascaded (MAGIC) pointing. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 246–253, 1999.

[121] Q. Zhang, A. Imamiya, K. Go, and X. Mao. Overriding errors in a speech and gaze multimodal architecture. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, 2004.