Paired Speech and Gesture Generation in Embodied Conversational Agents

by

Hao Yan

M.E., Electrical Engineering
Tsinghua University, 1998
B.E., Electrical Engineering
Tsinghua University, 1995

Submitted to the program in Media Arts and Sciences, School of Architecture and planning, in partial fulfillment of the requirements for the degree of Master of Science in Media Arts and Sciences at the Massachusetts Institute of Technology

June, 2000

© 2000 Massachusetts Institute of Technology. All rights reserved.

Author_____
Hao Yan
Program in Media Arts and Sciences
May 5, 2000

Certified by_____
Justine Cassell
Associate Professor of Media Arts and Sciences
AT&T Career Development Professor of Media Arts and Sciences

Accepted by_____
Stephen A. Benton
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Paired Speech and Gesture Generation in Embodied Conversational Agents

by

Hao Yan

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on May 5, 2000 in partial fulfillment of the requirements for the degree of Master of Science in Media Arts and Sciences at
The Massachusetts Institute of Technology

Abstract

Using face-to-face conversation as an interface metaphor, an embodied conversational agent is likely to be easier to use and learn than traditional graphical user interfaces. To make a believable agent that to some extent has the same social and conversational skills as humans do, the embodied conversational agent system must be able to deal with input of the user from different communication modalities such as speech and gesture, as well as generate appropriate behaviors for those communication modalities. In this thesis, I address the problem of paired speech and gesture generation in embodied conversational agents. I propose a real-time generation framework that is capable of generating a comprehensive description of communicative actions, including speech, gesture, and intonation, in the real-estate domain. The generation of speech, gesture, and intonation are based on the same underlying representation of real-estate properties, discourse information structure, intentional and attentional structures, and a mechanism to update the common ground between the user and the agent. Algorithms have been implemented to analyze the discourse information structure, contrast, and surprising semantic features, which together decide the intonation contour of the speech utterances and where gestures occur. I also investigate through a correlational study the role of communicative goals in determining the distribution of semantic features across speech and gesture modalities.

Thesis Advisor: Justine Cassell
Associate Professor of Media Arts and Sciences
AT&T Career Development Professor of Media Arts and Sciences

Paired Speech and Gesture Generation in Embodied Conversational Agents

by

Hao Yan

The following people served as readers for this thesis:

Reader_____

Alex Pentland
Professor of Media Arts and Sciences, Academic Head
MIT Media Laboratory

Reader_____

Matthew Stone
Assistant Professor of Computer Science
Rutgers University (New Brunswick)

Acknowledgements

I would like to thank my advisor, Professor Justine Cassell, for her continuous support and guidance during the whole thesis work and her push on me to meet all the deadlines. I would also like to thank Professor Matthew Stone for his generous help and patient explanation of his natural language generation system. Thanks to Professor Matthew Stone and Professor Alex Pentland for being my thesis readers and providing invaluable comments.

I would also like to acknowledge all other members in the REA project team. Without their diligent work and kind support, this thesis work would not be possible. Especially, thanks to:
Tim Bickmore for his great project management, insightful suggestions on implementation, valuable comments on my thesis, and generous CLIPS technical support.
Hannes Vilhjálmsson for his great recommendation on references, generous help on video equipments, valuable comments on my thesis, and always being willing to answer my questions.
Lee Campbell for his help on integrating the Festival speech system, easy-going ways, and always being willing to explain things to me.
Vikash Gilja for his help with transcribing part of the experiment data.
David Mellis, Ling Bao, and Nina Yu for their many contributions and pleasant company.

Studying and working in the GNL group has been such an enjoyable experience for me. I would especially like to thank:
Jennifer Smith for fighting along with me, generously sharing information with me, and always being kind to me.
Kimiko Ryokai for preparing me on how to deal with different kinds of people and also always being kind to me.
Mike Ananny for kindly reviewing my writings and giving invaluable feedback.
Petra Chong for always being easy-going and sharing job hunting experiences with me.
Andrew Donnelly for his diligent responsiveness to my many equipment and management requests.

Finally, I would like to dedicate this thesis to my mother, father, and all other family members. The values they instilled in me keep me pursuing higher goals. Their endless encouragement and comfort are my energy source for staying up at 3 o’clock in the morning.

Contents

1. Introduction

1.1 Scenario

Face-to-face conversation is about the exchange of information. In order for that exchange to proceed in an orderly and efficient way, participants engage in a series of elaborate social acts that involve communicative behaviors beyond mere exchanging of words.

In the following imaginary conversation, Tim and Lee were sitting in a meeting room waiting for a project meeting. They started a conversation about lodging in Boston area:

(1) Lee: …So, where do you live?
(2) Tim: I live in an apartment in Porter Square. It's about five minutes to the Star Market.
(3) Lee: How is your apartment?
(4) Tim: It's a beautiful place. I like it very much.

In this short exchange, many communicative behaviors were employed in addition to the actual speech, either for adding more information or for disambiguation. For example, in the second utterance of (2), by making a walking gesture (i.e. two fingers pointing to the ground and moving back and forth in opposite directions), Tim made it clear that his apartment is five minutes walking distance from the star market. The "on foot" information is conveyed through his gesture instead of speech. In (4), Tim makes an expansive gesture with his hands as a metaphor of "beauty". Meanwhile, the variation of intonation in the production of speech indicates the structure of this dialogue's propositional content. For example, in the first utterances of (2) and (4) there are pitch accents on "apartment" and "beautiful" respectively, indicating the new and most prominent information conveyed by the utterances. This spontaneous performance, which seamlessly integrates a number of modalities, is given unconsciously and without much effort[1].

1.2 Motivation

Today, the metaphor of face-to-face conversation has been increasingly applied to the design of human-computer interface. It has been shown that in many ways people treat computers as human

[1] See evidences reviewed in Cassell et al., 1999.

-- interaction with computers inevitably evokes human social responses (Nass et al, 1994). Early researches have demonstrated the promise of animated characters as human computer interfaces. For instance, many users consider presentations given by animated characters as being more lively and engaging (André, Rist and Mueller, 1998; Lester et al, 1997).

Interfaces that are based on face-to-face conversation may allow humans to communicate with computers naturally and easily, because humans have long years of practicing communication with other humans, and thus they need little training before they use such a system. An instance of this kind of interface is the embodied conversational agent, which is represented by a lifelike human or animal character and is capable of performing believable actions and reacting to human users. Adding a body enables multiple natural modalities of communication, and therefore might improve the robustness and effectiveness of the human-computer interaction, just as in the real-world human-human interaction.

In order for an embodied conversational agent to have a conversation with the user similar to the one that Tim and Lee engaged in, the agent should be able to recognize the user's verbal and non-verbal behaviors and give back in real time the same kind of response that humans give in a face-to-face conversation. Therefore, it is important to have a generation process in the system, which can form appropriate communicative behaviors based on the system's internal representation of its knowledge, understanding of the user's intention, and discourse context.

Studies have shown that non-verbal behaviors, especially gesture, play an important role in face-to-face conversation. An experiment by Nobe et al. (1998), in which gaze tracking methods are employed to track the user's attention when interacting with an anthropomorphic agent, shows that users do pay attention to the agent's hand gestures, especially to highly informative gestures or parts of gestures. Other studies, such as (McNeill, 1992) and (Cassell, McNeill, and McCullough, 1999), have shown that both speech and gesture can carry information, they appear to share the same underlying representation of linguistic and non-linguistic information, and that people do pick up the information conveyed by gesture during a conversation. Furthermore, while many nonverbal behaviors function at the turn level in discourse, such as gaze, brow movement, and body posture (Chovil 1992; Rosenfeld 1987; Laver 1975; Cassell, Torres, and Prevost, 1999), gesture and intonation are more likely to happen at the clause or phrase level and contribute to the content exchange in the discourse (McNeil 1992; Kendon, 1994; Cassell et al., 1994, Torres,

1997; Hirschberg, 1990). All these evidences suggest that speech, intonation, and gesture should be generated together, using the same underlying knowledge source.

However, little effort has been made to generate gesture in interactive systems. Even less effort has been made to generate gesture and speech together based on the same underlying representation. In this thesis, I pair the generation of gesture with the generation of speech, which will give us ample opportunity to further our understanding of the semantic relationship between speech and gesture, the synchrony of speech and gesture, and the role of communicative goals.

Moreover, once we admit that information can be conveyed by behaviors that use different communication modalities, a natural question is how to determine when to use which modality. The generation process in embodied conversational agents should be able to coordinate different modalities, most importantly speech and gesture, i.e., it should be able to decide when gesture occurs along with speech, what information gets displayed through gesture, how gesture is performed, and when a word or phrase in speech gets accented. Furthermore, it is essential for such a process to be able to decide when to use gesture to convey information that is complementary to speech and when to use gesture to convey the information that is redundant to speech.

1.3 Overview of Research Work

The ultimate goal of this thesis work is to provide in an Embodied Conversational Agent a framework that can generate both speech and gesture based on the same underlying representation of domain knowledge and discourse context in real time. In addition to building the generation framework, two particular multi-modal generation problems are explored in this thesis: distribution of the communicative load across modalities and intonation contour determination.

Gestures do not always carry the same meaning as speech. Those gestures that carry information not present in the simultaneous speech are called complementary gestures, while those that convey the same information as speech are called redundant gestures. One of the first decisions to make in the generation process is, what information should speech and gesture convey respectively, i.e. how to distribute the system's communicative load appropriately across speech and gesture modalities. A study of human-human conversation was conducted and correlation

between communicative goals, discourse information structure, semantics, and pragmatics was examined, in order to find rules that human interlocutors employ to distribute communicative loads based on intentions. These rules were then integrated into our generation framework for producing appropriate complementary or redundant gestures.

Intonation in speech is crucial in conveying the intended meaning in the discourse. A speaker chooses a particular tune, or intonation contour, to convey a particular relationship between an utterance, currently perceived beliefs of the hearer(s) and anticipated contributions of subsequent utterances (Pierrehumbert and Hirschberg, 1990). In a conversational system, inappropriate intonation selection can be seriously misleading and detract from the intelligibility of what is said. In this thesis, I also look into using discourse information structure, intentional and attentional structures, and contrast to determine intonation contours of speech utterances, as well as integrating this process with our gesture generation framework.

The thesis work is done in the context of the research on an embodied conversational agent. The agent, named REA (Real Estate Agent), plays the role of a real estate salesperson that interacts with users to determine their needs, shows them around virtual properties, and attempts to sell them a house. It has a fully articulated 3D graphical body and communicates using both verbal and non-verbal modalities. She is able to describe features of a house using a combination of speech utterances and gestures, and can also respond to users' verbal (via speech recognition) and non-verbal input (via computer vision). The system uses the SPUD (Sentence Planning Using Description) natural language generation engine (Stone & Doran, 1997) to carry out the response generation task. Figure 1.1 shows a picture of a user interacting with Rea.

Figure 1.1: User Interacting with Rea

1.4 Contribution

My major contribution in this thesis work is integrating speech, gesture, and intonation generation into a real time generation framework for embodied conversational agents that is capable of generating behaviors in both modalities based on the propositional representation of real-estate properties, discourse information structure, intentional and attentional structures (Grosz and Sidner, 1986), and a mechanism to update the common ground (or mutual beliefs, Clark and Marshall, 1981) between the user and the agent. Algorithms are implemented to analyze the discourse information structure, contrastive information, and surprising semantic features, which together decide the intonation contour of the generated speech utterances and where gestures occur. I also investigate through an empirical study the rules of distributing communicative load across modalities and the realization of multiple communicative goals in a single utterance.

This research has both theoretical and practical value. From a research point of view, there are many facets in face-to-face communication, such as content delivery, turn taking, requesting/giving feedback, etc. There are also different theories about the cognitive processes in a face-to-face conversation. By applying those theories to the generation process and observing the output behavior of our agent, we will be able to evaluate those theories. Moreover, it is also a good chance to examine some computational linguistic theories. For example, there are different theories about how to generate a sentence (or sentences) based on multiple communicative goals. Most of them still stay at the theoretical level. If we can realize those theories in our speech and gesture generation process, we can justify their feasibility. From an interaction design point of view, it is better to have system responses generated based on underlying knowledge and context, rather than canned in some sort of templates, because real-time generation provides more flexibility and variety in the response. For example, to refer to the same internal object ROOM1, the system could generate "a room", "the room" or pronoun "it", with or without an accompanying gesture depicting the sahpe of the object, depending on the salient information about the object in current context.

1.5 What the Thesis Is Not About

The gestures that I am interested in and tend to generate in this work are naturally occurring gestures in face-to-face conversation, or spontaneous gestures. They are in contrast to those intentional gestures, such as emblematic “ok” gestures or deliberate deictic (pointing) gestures

such as those used in many interactional system that use gesture as a equivalent means of input device (e.g. the “put-that-there” system by (Bolt, 1980). McNeill (1992) divides spontaneous gestures into four categories: iconic, metaphoric, deictic, and beat. Iconic gestures describe features of an action or event. Metaphoric gestures describe features of a concrete object to represent an abstract concept. Deictic gestures locate in space discourse entities with or without a physical instantiation. Beat gestures are abstract visual indicators that emphasize discourse-oriented functions. Among the four, iconic and metaphoric gestures are explored in the thesis.

Although natural language generation (NLG) is essential to the response generation in embodied conversational agents, this thesis is not trying to build an NLG system. Rather, I am more interested in the relationship between speech and gesture, the role of discourse structure in specifying verbal and non-verbal behaviors, and the representation and organization of the system’s internal knowledge. For the purpose of language generation, we use the SPUD (Sentence Planning Using Description) system, developed by Stone & Doran (1997).

1.6 Thesis Layout

This chapter introduces the motivation of the research and gives an overview of the research work. The rest of the thesis is divided into six chapters. The following chapter reviews the background of the thesis work, including the research in face-to-face communication, gesture, intonation, and previous interactive systems that deal with speech and gesture. Chapter 3 describes the semantic distribution problem, the experimental study on speech and gesture, and the results of the study. Chapter 4 discusses the role of intonation in embodied conversational agent systems and the implementation of intonation in the paired speech and gesture generation framework. Chapter 5 presents the implementation of the paired speech and gesture generation in the REA system, gives two detailed examples of generation, and evaluates the implementation. The following two chapters conclude the thesis and propose some interesting future work.

2. Context of Work

2.1 Face-to-Face Communication

Human face-to-face conversation is a multi-modal interaction. People get information not only from spoken words and variation of word intonation, but also from other non-verbal behaviors such as body posture, head nod, gesture, facial expression, etc. Many communicative behaviors have an interactional function (Goodwin 1981; Kendon, 1990). They serve, for instance, the purpose of regulating turns, and provide feedback cues that indicate the state of conversation. Others have a propositional function that complements or elaborates upon the information exchange in the communication. This thesis will focus on generating speech and gestures that have propositional functions.

Interestingly, the exchange of both interactional and propositional information in face-to-face conversation can be achieved by both verbal and nonverbal modalities (Cassell et al., 1999). For example, a speaker can request the turn by explicitly saying “ok, my turn.” She can also just move her hands into the gesture space to request the turn. Likewise, a speaker can use gestures to convey the “walking distance” information by doing a walking gesture while saying “It is 5 minutes to the star market.” She can also just say “It is 5 minutes walking distance to the star market.” Therefore, when developing an embodied conversational agent, which uses face-to-face conversation as an interface metaphor, it is appropriate to concentrate on the discourse functions of those behaviors, instead of the behaviors themselves. This principle leaves us two open-ended questions: (1) how to map behaviors from different modalities into unique discourse functions, and (2) how to realize a discourse function into a communicative action that involves appropriate speech utterance and non-verbal behaviors at the right time.

2.2 Relationship Between Speech and Gesture

Gesture is an integral part of discourse (Kendon, 1972; McNeil, 1992). Research has shown strong evidence that there is a close relationship between speech and spontaneous gestures in face-to-face conversations. More than three quarters of all clauses in naturally occurring narrative discourse are accompanied by gestures of one kind or another (McNeill, 1992). Especially when speech is ambiguous or in a noisy environment, people tend to produce more gestures and rely more on gestural cues for understanding (Rogers, 1978). Moreover, spontaneous gestures are

synchronized with speech. The most effortful part of a spontaneous gesture tends to co-occur or occur just before the phonologically most prominent syllable of the accompanying speech (Kendon, 1994). At the semantic and pragmatic level, gestures are also closely related to the accompanying speech. The two communicative channels never convey conflicting information, although they don't always carry the same information. Those concepts that are difficult to express in language, such as manner of actions and simultaneity of two events, may be conveyed by gestures (Kendon, 1994).

Several theories have been developed about the conceptual cause of the close relationship between speech and gesture. Butterworth and Hadar (1989) claim that speech is primary and gesture is a late occurring add-on to language, therefore gesture is not integral to communication. Some claim that instead of being used to communicate information, gesture is the encoding of information in the speaker's mind (Freedman, 1972). McNeill (1992) claims that speech and gesture are both integral parts of face-to-face conversation. They both arise simultaneously from an underlying representation that has both linguistic and visual aspects. This theory explains the strong temporal synchronization constraint on the production of both gesture and speech. It is also suitable to be used as the basis of computational realization of the parallel semantic and pragmatic content in speech and gesture. A speech-gesture mismatch study conducted by (Cassell, McNeill, and McCullough, 1999) found that listeners in conversation do attend to information conveyed in gesture and can retell the information in speech, even though they do not remember from which modalities they get the information. This result provides further evidence for McNeill's theory.

Gestures do not always carry the same meaning as speech. Those gestures that carry information not present in the simultaneous speech are called complementary gestures, while those that convey the same information as speech are called redundant gestures. One of the first decisions to be made in the generation process is to decide what information should speech and gesture convey respectively, i.e. how to distribute the system's communicative load appropriately into speech and gesture modalities.

Little research has been conducted concerning the above problem. Cassell and Prevost used information structure to predict when redundant and complementary information is conveyed across speech and gesture (Cassell & Prevost, 1996). Information structure describes the relation between the content of utterances and the emerging discourse context. The part of an utterance

that links to the previous utterances is called theme (or topic). The part of an utterance that forms the core contribution of the utterance to the discourse is called rheme (or comment). Cassell and Prevost proposed three rules:

(1) Rhematic information with a focus marking newness indicates complementary information across both modalities.
(2) Rhematic information with a focus marking contrast indicates redundant information realized in both modalities.
(3) Thematic information with a focus marking contrast indicates redundant information expressed by both modalities.

However, they did not explicitly explore the semantic relationship between speech and gesture. Torres made good efforts towards formalizing the condition of selecting semantic information to be expressed by gestures. He employed the concept of semantic structure (Jackendoff, 1983) and proposed three hypotheses (Torres, 1997):

(1) An implicit argument in a conceptual structure may serve as a representation of what is distributed in gesture as complementary information.
(2) A spatial structure encoding linked to a conceptual structure of a natural action may serve as a representation of what is distributed in gesture as complementary information.
(3) A marked feature in a conceptual structure may serve as a representation of what is distributed in gesture as redundant information.

The importance of these works for generating action descriptions is clear, but they have not been deployed in actual speech and gesture generation system.

A parallel kind of research lies in the area of multimedia presentation, where there is a need to distribute information to be described by text and by graphics form. Green (Green et al, 1998) developed systems that automatically generate presentations consisting of coordinated text and information graphics. Similar systems include the WIP project (Wahlster et al., 1991) and the COMET project (Feiner & Mckeown, 1990). The media allocation problem (when to use text and when to use graphics) in these systems is solved by applying rules about the kinds of information that text and graphics are good at expressing. While such methods are certainly enlightening for our research in semantic distribution across speech and gesture modalities, more information

from other aspects are needed for generating speech and gesture, due to the close relationship between speech and gesture, the great variety of gesture, and different overall communicative goals.

2.3 Intonation In Conversation

Intonation is crucial in conveying intended meaning in conversational discourse. In a conversational system, inappropriate intonation selections can be seriously misleading and detract from the intelligibility of what is said. For example:

Q1: I am looking for a condo in Boston.
A1: I *have* a condo in Boston.
Q2: I am looking for a place in Boston.
A2: I have a *condo* in Boston.

Answers A1 and A2 are the same sentence. However, by placing the pitch accent on different words, the speaker is able to answer two distinct open questions -- whether the addressee has any place in Boston and what kind of place the addressee has in Boston. Switching answers A1 and A2 in the two situations would not be appropriate.

Pierrehumbert and Hirschberg (1990) present a system of intonational descriptions that distinguishes accent placement (stress), tune, phrasing, and pitch range. Tunes are described as sequences of low (L) and high (H) tones that determine the shape of the fundamental frequency (f_0) contour. There are six types of pitch accents in English:

(1) H* - the most common pitch accent, it comes out as a peak on accented syllable.
(2) L* - phonetically realized as local f_0 minima.
(3) L + H* - pronounced with a valley followed by an accented peak, in which the tune starts above the f_0 baseline and climbs to a relatively long rest on the accented peak, on the second syllable.
(4) L* + H -pronounced as a longer sustained, accented valley that builds to a non-accented, narrow peak.
(5) H* + L -begins with a peak accent on the first syllable and descends in tune.
(6) H + L* -begins at a peak but descends quickly, with the accent on the "valley" syllable.

Research also shows that the intonation contour coincides with discourse information structures (Pierrehumbert and Hirschberg, 1990; Prevost & Steedman 1994; Steedman, 1999). In particular, the L+H* accent is associated with the theme (or topic) of an utterance which links the part of utterance to prior utterances, while the H* or H*+L accent is associated with the rheme (or comments), which forms the core contribution to the discourse (i.e. the new or particulary salient information). Moreover, Pierrehumbert and Hirschberg (1990) point out that speakers use tune to specify mutual beliefs (Clark and Marshall, 1981) of participants in the current discourse. In particular, phrases whose accents are all H* appear to signal to hearer that an open expression is to be instantiated by the accented items and the instantiated proposition realized by the phrase is to be added to hearer's mutual belief space.

The Previous Mention Strategy (Hirschberg 1990; Monaghan 1991) is employed by most text-to-speech systems to select an intonation pattern for a given speech utterance. It predicts pitch accents by a set of heuristics based on textual givenness, which is determined by searching the current word in a history list. Other researchers looked into modeling the intonational phenomena associated with semantic contrasts and specify accentual patterns based on sets of alternative properties from a knowledge base and a contrastive stress algorithm (Prevost 1996; Hiyakumoto et al., 1997). Currently, even the most sophisticated text-to-speech synthesis systems, cannot automatically generate appropriate pitch accent and contrastive stress patterns from given text, though many of them, such as the Festival speech synthesis system (Black and Taylor, 1997), support the realization of those intonation patterns based on given f_0 annotations.

2.4 Previous Systems that Generate Speech and Gesture

A few attempts have been made to generate speech and/or gestures in interactive systems. They are complements to earlier works that integrate speech and gesture input such as Put-that-there (Bolt, 1980). Rijpkema and Girard (1991) generated handshapes automatically for an animated character based on the object being gripped in the scene. Perlin and Goldberg (1996) employed rhythmic and stochastic noise functions in developing a system that allows real-time generation of lifelike behaviors for animated actors. In an effort to build “virtual human”, Badler et al. (Badler et al., 1999) presents an architecture for real-time motion and language-based avatar control interfaces. They developed a Parameterized Action Representation to be the intermediate

structure between natural language instructions with complex semantics and task execution by a virtual human agent.

André and Rist (2000) built a system for generating conversation between multiple lifelike characters for the purpose of presenting information to a human, in the domain of car sales and soccer commentary. The system automatically generates the characters' performance, which is represented as a sequence of communicative acts and decomposed into smaller communicative acts. The modalities explored are primarily speech and intonation, although there are some deictic gestures. The speech utterances are generated using a template-based approach. In Cosmo, a pedagogical agent (Lester et al., 1999), Lester and his colleagues developed a spatial deixis framework, which can generate deictic gestures and corresponding referring expressions in speech based on the representational structures in a world model, a curriculum information network, a user model, the current problem state, and the focus histories for gesture and speech. In their framework, the generation of gesture and speech are accomplished in two separate processes. Similarly, Rickel and Johnson (1999) have their pedagogical agent move to objects in the virtual world that it inhabits, and then based on templates, generate a deictic gesture at the beginning of the verbal explanation that the agent provides about that object. However, in their system, gestures are always redundant, complementing the verbal explanation.

Absent from these systems is the consideration of discourse structures and pragmatics for specifying non-verbal functions and the ability to allocate communicative load to different modalities.

In Animated Conversation (Cassell et al., 1994), an interaction between two autonomous graphical agents was implemented. It was the first system to automatically produce context-appropriate gestures, facial movements, and intonational patterns for animated agents based on deep semantic representation of information. However, instead of specifying handshapes and hand movements according to semantics, gesture forms are selected from a canned gesture dictionary, which provide particular gesture forms for particular concepts in the domain of the dialogue. Moreover, the generation process could not run in real time.

3. Semantic Relationship Between Speech and Gesture

In many ways people treat computers as human. Interaction with computers often evokes human social responses (Nass et al., 1994). While using an embodied conversational agent as a human-computer interface might stimulate natural and robust interaction, it also raises the bar for the believability of the agent's behavior. Mistakes in the generation or production of an embodied conversational agent's output behavior, such as gestures occurring on wrong word, misplaced accent, etc., might cause the user to misunderstand, be frustrated, and thus doubt the believability of the agent's expertise. To avoid such mistakes, we need to base our embodied conversational system on actual studies of human face-to-face conversation. The system should be implemented based on the essential properties found in those studies.

For this purpose, we conducted an empirical study about speech and gesture relationships in human face-to-face conversation. In the study, we are especially interested in the correlation among communicative goals, semantics, and the distribution of semantics into speech and gesture modalities.

3.1 House Description Experiment

The data of this study were naturally occurring narrative discourse collected from a house description experiment, in which human subjects pair up and describe a particular house. At the beginning of each experiment, a subject (the narrator) was asked to study a floor plan and a 15-minute video that walks through a house. The subject can watch the video and study the floor plan in any way they like. After that, the subject was asked to describe the house to another subject (the listener) who was unfamiliar with the house and had not seen the video. The narrators were told to describe the house in a way such that the listener could describe back the salient features of the house. During the conversation, the floor plan and the video were not available to either of the subjects. The listener was allowed to interrupt and ask questions whenever the narration was not clear. There are several benefits of the specific experiment setup:

(1) By providing materials about a particular house, we can naturally constrain the conversation in a certain domain instead of random topics. Thus we are able to observe

across different conversations on the same topic and find out common behaviors people have.

(2) The experiment domain is chosen to be house description because spatial description is more likely to stimulate naturally occurring gestures. Moreover, it is close to the domain of our system (real estate), thus the results found here would be directly applicable to the system.

(3) Since we know what the subject is describing, it would be easy to judge if the subjects gestures carry meanings and if the meanings carried by gestures are complementary to speech.

The narrations were videotaped using a video camera placed in a distance so that it is likely to be ignorable to the subjects and the narrators' upper body space were completely visible and their voices could be comfortably heard. The subjects were comfortably seated facing to each other in unarmed chairs, with ample free space in between so that they could move their hands without obstructions. All the equipments were set up and started before the subjects started their conversations and were not stopped until the conversations ended. The whole narrations of the narrators were videotaped without altering the focus, zooming in or out, or increasing or decreasing the level of sound. Figure 3.1 shows a video frame of a subject while uttering "There is a staircase in the middle (of the house)".

All participants were given the same instructions in the experiments. They were told that the data collection was part of a research effort to study language use patterns in face-to-face conversations, without mentioning that gestures were part of the primary areas of interests. All subjects, recruited from MIT campus, were adult native speakers of American English. Ten narrators, five of them male and five of them female, were video taped with their consent. The length of conversations ranges from 10 to 30 minutes.

3.2 Transcription And Analysis Procedure

Out of the ten conversations that were videotaped, five that have better sound quality have been transcribed. Speech of both interlocutors is fully transcribed, with each utterance being time-stamped at the beginning and the end. Those gestures that carry semantic information, no matter redundant or complementary, are also recorded. These include what McNeil refers to as representational gestures (iconic and metaphoric) that represent attributes, actions, or

relationships of object or characters (McNeil, 1992), and deictic gestures that are finger points or other indications of either concrete or imaginary objects. In the rest of the thesis, I refer to these gestures as referential gestures. Beats are not recorded and analyzed, because they are more like indicators of discourse structure than carriers of semantic meanings. The generation of beats in our generation framework is based on studies such as (McNeil, 1992) and (Cassell et al., 1994), which show that beats tend to co-occur with the rhematic part of speech utterances.

Figure 3.1: “There is a staircase in the middle.”

Gestures are transcribed following McNeil’s coding scheme for iconic gestures (McNeil, 1992). In accordance with our implementation of gesture generation (which synthesizes gesture from basic gesture elements), we record basic aspects of each gesture, which include: 1) which hand (s) is used; 2) the hand shape (s); 3) the shape of trajectory; 4) and the starting position of trajectory. Spatial positions of gestures are specified with reference to zone divisions. The three-dimensional gesture space in front of the speaker can be divided into a series of concentric squares. The square that is directly in front of the chest is Center-Center. Surrounding it is the Center, and then Periphery. Finally, the Extreme Periphery represents the rest of the outer square. Each concentric square is further divided into upper, lower, left and right. Figure 3.2 shows the division of the gesture space. A position in the gesture space therefore can be represented by a combination of those specifications, such as “periphery upper left”. Most hand shapes are specified according to ASL (American Sign Language) hand shape specifications, such as L, 5, O, and G. To facilitate the later understanding of the gesture from its transcription during the analysis procedure, we used natural-language-like descriptions instead of using acronyms and abbreviations. Below is the transcription of the gesture shown in figure 3.1, which shows complementary information that the

staircase is spiral shaped: “Right hand loose H shape, moving spirally up from periphery center right to periphery upper right”.

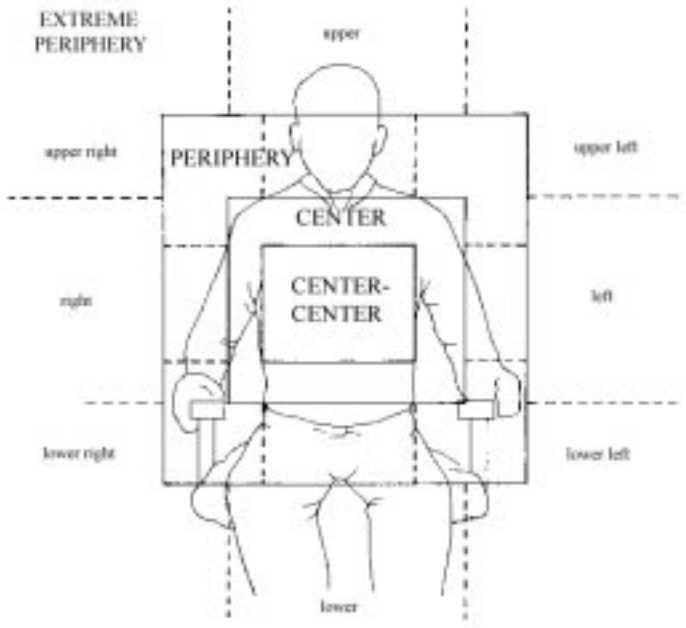


Figure 3.2: Division of the gesture space. From McNeil (1992).

The primary purpose of this experimental study is to look at the role of communicative goals in determining the semantic distribution across modalities. Therefore it is important to encode semantic features conveyed by speaker’s speech and gesture and the goals the speaker wants to achieve by the whole communicative action. Moreover, to facilitate the correlational analysis, it is important to have a finite number of semantic features. Therefore, instead of recording detailed semantic meanings such as “left”, “right”, “square”, and “circle”, we record semantic categories such as “Location” and “Shape”. Jackendoff (1983) proposed the semantic (conceptual) structure, which assumes that meanings are mentally encoded and decomposable. He argues that there are several essential units of conceptual constituents in conceptual structures. These conceptual constituents belong to ontological categories such as THING, EVENT, STATE, PLACE, and PATH. Inspired by this claim and based on the domain of the house description experiment, we developed an extended list of semantic categories that can be used in the semantic coding procedure. The list adequately covers the semantic features in most gestures and speech that we looked at. The semantic categories we used are:

Shape: the shape of an object, e.g. “rectangle”, “square”.

Size:	the size of an object, e.g. “big”, “small”, “about 15 square feet”.
Location:	the location of an object, e.g. “The living room is to the right.”
Relative Position:	semantic information about the position of an object relative to other objects, e.g. “If the living room is here, the library would be here” (gestures show the relative position information).
Existence:	the existence of an object, e.g. “There is a kitchen.”
Action:	information about an action, e.g. “put”, “go”.
Path:	the path or trajectory of a motion.
Direction:	the direction of a motion (action), e.g. “Then, you go to the right.”
Manner:	the way to perform an action, e.g. “walk” is a manner of “go”.
Impression:	general impression about an object (or objects), e.g. “luminous”, “great”.

For each utterance accompanied with gesture of our interest, we code categories of semantic features that are conveyed by speech and those that are expressed by gestures. If the semantic features expressed by gesture overlaps with those in speech, the gesture is classified as a redundant gesture. On the other hand, if the semantic features in gesture cannot be found in the semantic features conveyed by speech, the gesture is classified as complementary gesture. When judging the complement or redundancy of a gesture, the actual semantic meanings are also taken into account besides semantic categories. So if there is the same semantic feature in both modalities but the corresponding semantic meanings are different, the gesture is considered as complementary. For example,

Speech: It (the first floor bathroom) is adjacent to the living room.

Gesture: hands show that the bathroom is to the left of the living room.

In the above example, both speech and gesture convey the Relative Position information. However, the “to the left” information is only found in the gesture.

Based on the semantic features conveyed by both speech and gesture, the current context such as the experimenter’s open questions, and the actual features of the object that the speaker is describing, we determine the speaker’s main goals of the whole communicative behavior. The communicative goals are similar to those semantic categories except that they are in the form of “describe/explain Shape”. For simplification, we omit the “convey/explain” part and use similar ontology as those semantic categories to define communicative goals. The communicative goals defined and used in the analysis process are:

Location: describe the position of an object, e.g. “The bedroom is to the right.”

Introduction: introduce a new object or objects, e.g. “There is a staircase in the middle of the house.”

Configuration: describe the placement of several objects, e.g. “The living room is here and the library is here.” (gestures show the relative position)

Action: describe an action, e.g. “you put them in.”

Impression: describe a general impression about an object, e.g. “The bathtub is cool!”

Shape: describe the shape of an object, e.g. “The ceilings are concave.”

Enumeration: enumerate number of objects, e.g. “It has three stories.”

Most of the utterances in the transcription can be fitted into those categories. There are 7 out of the total 328 utterances that cannot be appropriately classified into any of the above categories, which we code as “Other”.

Of the 328 utterances transcribed, 244 (74.4%) were analyzed by two people in parallel. The inter-reader reliability (the percentage of analysis on which the two people agreed) is 85.7%. 89 out of 134 referential gestures are coded in parallel by the same two people. The inter-reader reliability is 97.8%. The final coding used for the analysis is agreed by both people.

Statistics are then made to study the frequency of the appearance of each semantic category in different modalities and in difference situations, and the relationship between such distribution and communicative goals. Appendix A shows a sample piece of transcription and analysis of the experimental data.

3.3 Result – Semantic Features in Gesture

A total of 328 utterances are observed in the analysis of 5 naturally occurring face-to-face conversations. 40.9% (134) of these utterances are accompanied with referential gestures. Interestingly, approximately half of these gestures (70) convey redundant information to speech, while another half contains information that is complementary to the accompanying speech. We did not record beat gestures, although we noticed that most utterances, especially those utterances without referential gestures, are accompanied with one or more beats. These facts indicate that gesture plays an important role in the face-to-face communication, and gesture does have the

function of conveying meanings. Most (127) of the 134 gestures are iconic. There are also a few metaphoric (8) and deictic (7) gestures.

Semantic Features	Semantic Features in All Gestures	Semantic Features in Complementary Gestures	Semantic Features in Redundant Gestures
Shape	33	23	10
Size	5	0	5
Location	31	9	22
Relative Position	28	26	2
Existence	0	0	0
Action	9	0	9
Path	13	11	2
Direction	9	2	7
Manner	1	1	0
Impression	3	0	3
Number	3	0	3
Other	1	0	1

Table 3.1: occurrence of semantic features in gestures

In terms of the actual semantic meanings carried by gestures, they are grouped into a few semantic categories. Table 3.1 summarizes the number of times that different semantic features occur in all gestures, complementary gestures, and redundant gestures. In detail, figure 3.3 shows the distribution of semantic features among all the 134 gestures studied. The "Other" category in the figure actually combines those low frequency semantic categories in table 3.1, including Manner, Impression, Number, and Other. Those semantic features most frequently appeared in gestures are: Shape, Location, Relative Position, and Path. This result may help us understand which semantics is more likely to be realized in the gesture channel in the generation process.

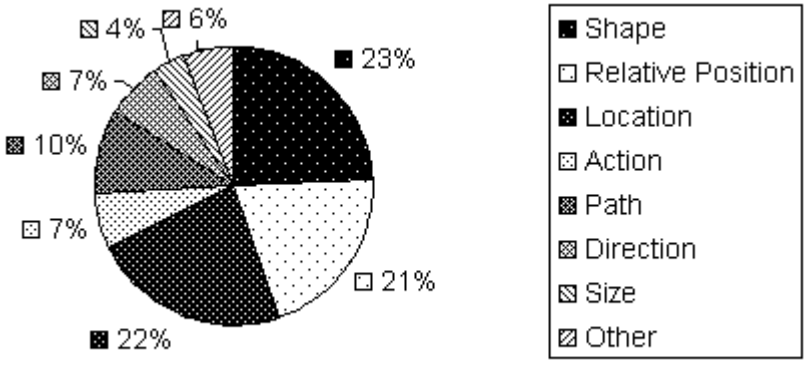


Figure 3.3: Semantic features in all gestures

Further, there is a clear distinction between the distribution of semantic categories in complementary gestures and in redundant gestures. Figure 3.4 shows the distribution in the 70 complementary gestures. It shows that only a few semantic categories are likely to appear in complementary gestures, such as Shape, Relative Position, Path, and Direction. The percentages of these categories are significantly higher than others.

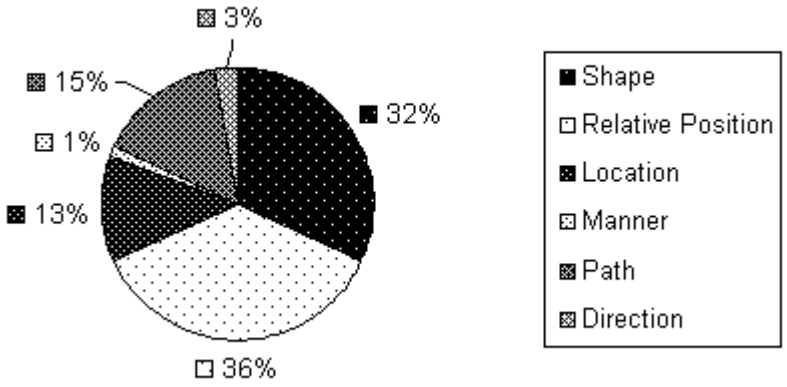


Figure 3.4: Semantic features in complementary gestures

On the other hand, in redundant gestures, more semantic categories are likely to appear (see figure 3.5). The Location feature is significantly more frequent than other categories and the distribution of the rest semantic categories are more averaged. These results may help us understand when is appropriate to add new information in the gesture channel in the generation process.

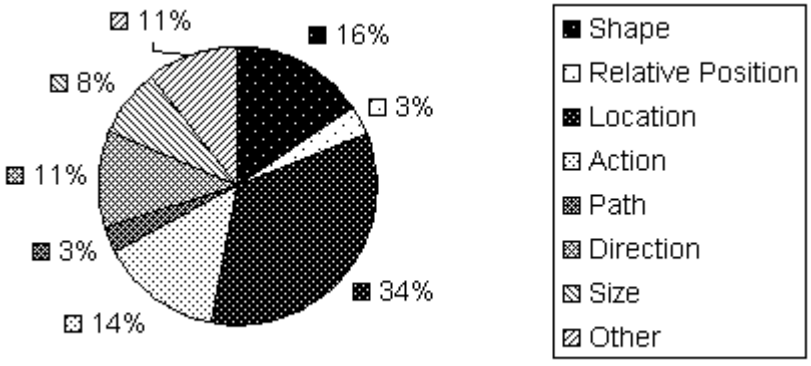


Figure 3.5: Semantic features in redundant gestures

Notice that a gesture can carry more than one semantic feature and many gestures that are classified as complementary gestures also carries redundant information. Figure 3.4 takes into account the redundant semantic features from this kind of gestures.

Moreover, these distribution results are largely dependant on the conversation domain. For example, in another study that looks at action descriptions (Cassell and Prevost, 1996), Manner and Path are found to be the semantic features that most frequently appear in complementary gestures. However, these results in different domains are not contradictory to each other, since goals that the speakers want to achieve are different in different domains. This fact leads us to study another dimension of the communication, which is communicative goals, in the decision of semantic distribution across modalities.

3.4 Result – Semantic Features and Communicative Goals

The communicative goal is the goal a speaker wants to achieve by performing a communicative action such as a speech utterance accompanied by a gesture. It is defined in terms of what the speaker wants the hearer to believe after hearing (seeing) her communicative action. In our data analysis, we found that communicative goal may constrain or decide the situation of semantic distribution across modalities.

The analysis method that we used in this study is similar to the method that many researchers use to find rules of predicting intonation contours for speech (Brown, 1983, Altenberg, 1987, Hirschberg 1990, Steedman and Prevost, 1994, etc.) In the intonation research, people study the correlation between intonation contour and the syntactic or semantic structure of the speech. Then the principles found in the study are used to predict intonation contour given a particular syntactic or semantic structure. In our study, we first code the major communicative goals in each utterance that has referential gestures. We then investigate the connection between communicative goals and the type of gestures (complementary or redundant), and further the connection between communicative goals and patterns of semantic distribution across modalities. We hope to generalize these correlations into heuristics of determining semantic distribution across speech and gesture modalities, i.e., given a particular communicative goal, the heuristics decide the semantic features to be realized in speech and gesture respectively. Table 3.2 summarizes some heuristics of determining the semantic distribution that we found in the house description domain. Each row in the table stands for a heuristic. The heuristic is a function of the major

communicative goal. The last column in the table gives the accuracy of those heuristics, i.e. percentage of occurrences of utterances that have the same major communicative goal and support the heuristics. A few of these heuristics in the table are described in detail in the rest of this section.

Major Communicative Goal	Gesture Type	Semantic Features in Speech	Semantic Features in Gestures	Accuracy in Data
Introduce a single Object	Complementary	Existence	Shape, Location	78.8% (26/33)
Introduce multiple Objects	Complementary	Existence, Number	Relative Position	100.0% (9/9)
Describe the configuration of multiple objects	Complementary	Existence	Relative Position, Shape	93.8% (15/16)
Describe location of an object	Redundant	Location	Location	80.0% (20/25)
Describe a general impression	Redundant (Metaphoric)	Impression	Impression	100.0% (6/6)
Describe the shape of an object	Redundant	Shape	Shape	100.0% (5/5)

Table 3.2: Communicative goals and semantic features

3.4.1 Introduction Heuristic 1 – Single Object

One effective way of describing a place is to introduce a new object in the place. Approximately 31.3% of gestures occur in those utterances where the communicative goal is to introduce object(s). The sentence structures that speakers choose to realize the introduction goals are mostly "There be" type and the gestures are mostly complementary. Especially,

(1) When introducing a single object, a complementary gesture accompanies the speech describing certain features of the object, mostly Shape, Location, and Relative Position (to other previously mentioned objects).

There are 33 instances in our data whose communicative goals are introducing single objects. 26 out of the 33 (78.8%) instances support heuristic (1). For example:

a. While speaking "There is a staircase", the speaker uses gesture to describe the spiral shape of the staircase.

b. While speaking “There was sort of a reading room”, the speaker moves her hand to the left showing that the reading room is off to the left side.

There are several possible reasons to explain why people use gesture instead of speech to express Shape or Location information. First of all, introducing certain semantic features is very important to achieving the overall goals of the communication in a certain domain. Therefore, people tend to introduce those semantic features whenever possible. In the house description domain, the Shape and Location are among those important features. Second, gesture is good for expressing those semantic features that is difficult to express in speech (Kendon, 1994), especially those spatial information like Shape and Location. For example, instead of speaking “There is a bathtub that is like a wide flat-bottomed vessel that has round corners”, people can just use their hands to depict the contour of the bathtub. The hearer in this situation would probably get a better understanding of the shape when she saw the gesture. Third, if there is some uncommon or surprising feature about the object that the speaker is introducing, to avoid misunderstanding, it is important to introduce it as soon as possible. Therefore the surprising semantic information could be expressed by modality whenever appropriate. The “uncommon” or “surprising” here means that the value of a semantic feature is not assumed in common sense. For example, a porch in common sense is a rectangle area in front of or behind a house, however the house in our experiment has a porch that wraps around it, so almost every subject in our experiment noticed that and produce a gesture to describe the wrap-around feature. Lastly, sometimes it is important to identify an object being described from its alternatives on a certain semantic feature, such as Shape. This can be done by performing a gesture that describes the semantic feature. For example, there are different kinds of tables, such as round table, rectangle table, or triangle table. To make sure that the hearer would understand exactly which kind of table is introduced, the speaker can use a gesture to depict the round shape of the table.

There are 7 out of the 33 (21.2%) instances in our data that do not support heuristic (1). 2 of the 7 exceptions have gestures that emphasize a certain semantic feature. For example, while speaking “The third floor has this **huge** living area”, the subject did an expansive gesture to emphasize the Size feature. In the rest cases, gesture still carries the same kinds of semantic features as those utterances that support the heuristics, such as Shape and Location, but those features are also expressed by speech. Moreover, gestures in these instances are pretty much lexicalized, i.e., the semantic features that those gestures carry are intrinsic meanings to some particular words, such as “flat” feature of the word “floor” or “steps” feature of a staircase. People would assume those

intrinsic semantic features when they hear those words. So the lexicalized gesture does not add any information. A possible reason for why people produce this kind of redundant gesture is that, these gestures mark some key features of the associating words, which are important to the understanding of the words.

3.4.2 Introduction Heuristic 2 – Multiple Objects

In the house description experiment, the Relative Position feature almost always appears in the gesture modality and complementary to the accompanying speech. Those complementary gestures usually happen when subjects introduce several objects (either the same kind or not) in the same utterance. Therefore we can predict complementary gestures when the major communicative goal is to introduce multiple objects:

(2) When introducing multiple objects, complementary gestures will probably occur with the speech describing the Relative Position (layout) of objects. The gestures are usually in the form of a series of gestures (either beats or gestures describing shape of object) in different space locations.

The "multiple objects" in the heuristic could be either a set of objects of the same type, such as "There are the cabinets", or objects of different types, such as "There is a living room, a library, and a dinning room on the first floor". There are 9 instances in our data that introduce multiple objects, all of which support this heuristic. For example: While speaking "The outside has like a porch going around, and then (there are) these columns" the speaker repeatedly use right hand to depict the shape of columns in different space positions.

3.4.3 Location Heuristic

In the above heuristics, the Location feature frequently appears in complementary gestures in different situations. However, when Location itself becomes the communicative goal, i.e. to explain the location of an object, a redundant gesture is more likely to occur, which repeat the location information. For example, while speaking "it (the living room) is off to the right", the speaker moves her right hand to the right. Moreover, the redundant gestures produced in this situation are lexicalized and associated with typical direction adverbs (right, left, up, down, etc.) and prepositions (up, underneath, etc.). There are 25 instances in the data whose main goal is to describe the location of an object. 20 out of the 25 instances (80.0%) support this heuristic.

(3) When describing the location of an object, a redundant gesture will probably occur, depicting the Location feature. The gesture is usually in the form of deictic, or lexicalized gestures.

There are 5 out of 25 instances (20.0%) in the data that do not support the location heuristic. 3 instances out of the 5 exceptions contain complementary gestures that introduce some uncommon semantic features of the objects being described. For example, while answering the question "where is the door (to the kitchen)?" the subject speaks "the door is on the porch" and gestures the uncommon shape of the porch. The other two exceptions contain complementary gestures that describe the relative position of the objects mentioned in the speech. For example, while speaking "The bathroom is between the living room and the library", the subject used two beats to show that the living room is to the right of the bathroom and the library is to the left. These exceptions, along with the introduction heuristics, suggest that complementary gesture has higher priority than redundant gesture to appear in the final communicative action.

3.5 Conclusion from Results

In all 328 utterances that we studied, the main communicative goals are always addressed directly by speech utterances, no matter what meanings the accompanying gestures would convey. Gesture either helps communicate the main communicative goal (such as identifying the position of an object in relation to another object when introducing the location of the object), or carry out other helpful information other than the main communicative goal (such as describing Shape information when introducing an object). Overall, gesture is produced to mark the key information that is crucial to the realization of the overall goal of the conversation in the specific domain – to walk the hearer through as many features about the house as possible. Complementary or redundant, gesture appears in the whole communicative action only in several situations:

(1) A beat gesture can occur in any utterance as an indicator of the information structure. Beats usually co-occur with the rhematic part of the accompanying speech[2].
(2) When there is a lexicalized gesture associated with a semantic feature and communicating the feature is important for achieving the overall goal of the communication, a redundant

[2] We did not transcribe and analyze beat gestures in our study. This result is based on Cassell et al, 1994.

gesture can be produced. This is based on the fact that 72.7% of all redundant gestures in our data are lexicalized. Although the information carried by this kind of gestures is redundant, performing these gestures might make the communication more vivid and robust.

(3) To achieve the overall goal of the conversation, there are often other communicative goals than the major communicative goal. We call those goals secondary goals. Examples of secondary goals include describing Shape or Relative Position when introducing an object or objects. Semantic features referred to by secondary goals can be expressed using a complementary.

(4) When there is some semantic feature that is uncommon (surprising) about the object being described and when it is appropriate to use gesture to describe that meaning, a complementary gesture will occur with the mentioning of the object in speech.

(5) When it is necessary to identify the object being described from its alternatives, a complementary gesture will occur with the mentioning of the object in speech. The importance of identifying a particular semantic feature is usually domain specific. For example, in the real estate domain, the location of an apartment is important. Therefore, it is valid to identify the height of a condo in an apartment building when introducing the condo.

3.6 Discussion of Implementation

From the rules summarized above, we can see there is a hierarchy for the realization of gestures, shown in figure 3.7. At the lowest level is the beat gesture. There is no specific constraint for generating a beat gesture except that it co-occurs with rheme. Therefore, in the implementation of the paired speech and gesture generation, beat can be realized as a default gesture at the rheme of the speech utterance. It will be overridden in some certain situations.

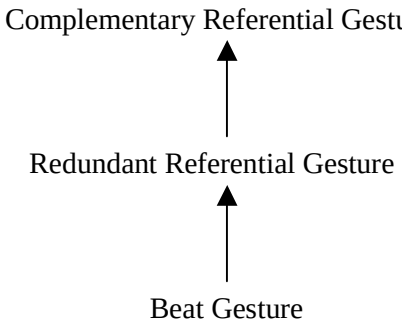


Figure 3.7: Gesture Hierarchy

The generation of other gestures, especially gestures that carry semantic meanings, would be constrainted in the context. The higher the gesture in the hierarchy, the more constraints would be applied for generating that gesture. The following pseudo code gives the gesture generation algorithm when the major communicative goal is introducing a object or many objects. Gesture generation for other communicative goals is similar.

```
If <surprising semantic feature X> then
        Generate complementary gesture about X
Else If <necessary to identify the object from its alternatives on semantic feature Y> then
        Generate complementary gesture about Y
Else If <secondary goal Z>
        Generate complementary gesture about Z
Else If <semantic feature R has lexicalized gesture and is important in domain> then
        Generate redundant gesture about R
Else
        Generate a beat
```

The constraints can be represented in the form of semantic propositions, pragmatics, organization of semantic propositions, and communicative goals. For example, the secondary goal can be represented by putting semantic propositions in the same category, so that when a proposition in that category become the main communicative goal, the rest propositions in the same category become the secondary goals; the appropriateness of using a gesture to express a semantic feature can be represented by rules that connect semantic meanings to gesture components; the domain-specific need for identifying an object from its alternatives can also be represented by grouping semantic propositions into the same category; the surprising semantic feature of an object can be represented by pragmatics. Chapter 6 discusses the detailed implementation of the paired speech and gesture generation.

4. Intonation Generation

In naturally occurring speech, some words or phrases are more intonationally prominent than others. This intonation change across words is different from lexical stress patterns associated with words. For the same utterance, the intonation pattern could be different when produced in a different context. It reflects the speaker's intentions by indicating propositional prominence within the utterance. Intonation is thus crucial in conveying intended meaning in conversational discourse. Inappropriate intonation selections can be seriously misleading and detract from the intelligibility of what is said. See the example discussed in Section 2.3.

In Embodied Conversational Agent systems, the right intonation variation in the agents' speech output is crucial for making the agent believable. Currently available text-to-speech synthesis systems lack the capability of correctly predicting the intonation contour for any given utterance, although some of them can interpret hand-coded intonation specifications to produce f_0 contour[3]. Moreover, speech (including intonation) and gesture are integral part of a larger communication action and their generation refers to the same underlying knowledge source such as discourse information structure. Therefore, the intonation generation should be integrated in the paired speech and gesture generation.

4.1 Predicting Intonation

Many researchers have studied the relationship between syntax of speech and intonation and between semantic information and intonation. Their results can in turn be used to predict the intonation contour of a given speech text using available syntactic, semantic, or even pragmatic information in discourse.

Researchers who look at syntactic connection to intonation have agreed that accenting and de-accenting of words in a sentence can be determined by a combination of word class (open and closed classes), syntactic constituency, and surface position (Brown, 1983, Altenberg, 1987, Hirschberg 1990, etc.). For example, open class items in a sentence, such as nouns and verbs, are usually accented, while closed class items, such as articles, prepositions, and pronouns, are usually de-accented. Most text-to-speech systems use this rule to determine default intonation

[3] For definition of the f_0 contour, refer to section 2.3.

contour of a given piece of text. Hirschberg (1990) looked at lexical givenness (Prince, 1981) and proposed the previous mention strategy, as follows:

a. Assign accents to open-class lexical items.
b. De-accent all closed-class lexical items.
c. De-accent any lexical items that were already mentioned in the local discourse segment.

Some researchers look beyond the syntax level, into semantic information such as information structure. Information structure refers to the packaging of information within an utterance. An utterance is divided into several semantic propositions (Prevost & Steedman, 1994). For example, the utterance A2 in the above example can be divided into:

$$\begin{cases} \lambda x.have(I,x) \\ condo(c) \wedge have(I,c) \end{cases}$$

These propositions are then classified as theme and rheme. The theme represents the topic of the utterance, such as the $\lambda x.have(I,x)$. It is a link to prior utterances. The rheme of an utterance provides the core contribution of the utterance, such as the $condo(c) \wedge have(I,c)$. Steedman and Prevost (Steedman and Prevost, 1994, Prevost, 1996) have argued that the information structure decides the placement of pitch accent and boundary tones. Specifically, the L+H* LH% marks the theme and the H* LL% marks the rheme. Applying this heuristics, the intonation contour of the A2 is:

(I have) (a condo in Boston).
H* LH% L+H* LL%

In our paired speech and generation framework, we incorporate a combination of the previous mention strategy and the information structure heuristics. The Previous Mention Strategy is used to generate the default intonation for all words in an utterance, while the information structure heuristics will override the default intonation for the key words in theme and rheme. The discourse manager in our generation framework keeps track of a list of entities mentioned before. Thus we can use the entity givenness to approximate lexical givenness used by the Previous Mention Strategy. The entity givenness, theme, and rheme are analyzed in the discourse manager and passed to the natural language generator as pragmatics. The entity givenness can be generated by comparing the current discourse entity to a history list of previously mentioned discourse entities. The information structure of generated utterances can be derived from the speech act to be realized, which is passed from the discourse manager to the generator. These speech acts are usually in the form of "SA_DESCRIBE OBJECT o ASPECT x" or "SA_OFFER OBJECT o". In

the first one, the combination of the OBJECT and the ASPECT define the theme, which is $\lambda y.x(o, y)$. Since in our system open questions are always general questions about some entity raised by a recent turn, we approximate this proposition to theme(o) for simplicity. In the latter case, the OBJECT is the rheme, represented as rheme(o). Since the theme and rheme are complementary to each other, it is sufficient to supply either theme or rheme.

4.2 Contrast and Intonation

While the combination of previous mention strategy and information structure heuristics can correctly predict the intonation contour for most utterance, there are still cases that it cannot cover. For example:

a. I like **blue** tiles more than **green** tiles.
b. I like blue **tiles** better than blue **wallpaper**.

In those two utterances, the bold-faced words are stressed. The two utterances have exactly the same syntactic structure, same theme, and similar rhemes. However, the intonation contours are quite different. This is because there is a stronger constraint than information structure – contrast. The two utterances contrast on different things: utterance a contrasts on colors of two objects while utterance b contrasts on the two objects. In a situation like this, i.e. utterance with contrast, the contrastive stress will override the existing intonation contours. The analysis of contrastive semantic features among entities is discussed in the following section.

Therefore, similar to gesture generation, intonation generation also relies on a hierarchy of heuristics. The heuristic at the bottom of the hierarchy is used by default. The higher the heuristic is in the hierarchy, the stronger the constraints on the intonation generation. Figure 4.1 shows the hierarchy.

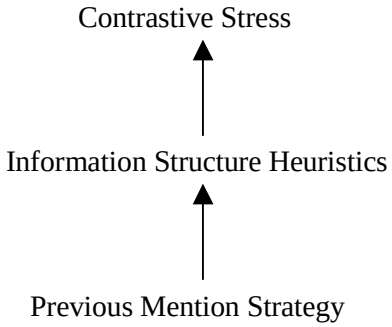


Figure 4.1: Intonation generation hierarchy

4.3 Implementation of Contrast Analysis

Prevost (1996) proposed a Contrastive Stress Algorithm, which analyzes the set of semantic features that discriminate a discourse entity from some other salient entities. His algorithm can be summarized as follows:

a. let RSET include x and its alternatives (determined by the previous discourse).
b. let PROPS be a list of all properties (features) of x, ordered so that nominal properties take precedence over adjectival properties.
c. let CSET be the (initially empty) set of properties of x that must be accented for contrastive purposes.
d. for each property p in PROPS, do following: if p is not a property of each member of RSET, eliminate those entities from RSET for which p does not hold and include p in CSET
e. stop when RSET contains only x.

Following this algorithm, we implemented the contrastive analysis in our paired speech and gesture generation system, which is able to analyze contrastive semantic features and generate pragmatic propositions in the form of "contrastive(Object1, Object2, Feature)". There are three kinds of knowledge bases that the contrastive analysis refers to. The first one is a sort of lexicon. It defines some basic syntactic constituents of an object and prominent semantic feature categories associated with the object. Each item in the "lexicon" is in the form:

(lexicon Name PartS Num Gender FeatureNameList...)

in which "lexicon" is a keyword; "Name" is the name of the object, or the kind of objects; "PartS" is part of speech; "Num" is number; "Gender" is gender of the name; "FeatureNameList" is a list of categories of semantic features that people usually associate with this object. Here is an example of an item in the "lexicon":

(lexicon room NP single neutral Size Location Light Style)

The second kind of knowledge base defines normal values of semantic features. Each item in this database simply lists all the values of features that are considered common. For instance, small,

medium, and large are considered common for Size, while huge is not. Below is a couple of examples of items in this kind of knowledge base:

(normalvalues Size small medium large)
(normalvalues Location left right up down)

The third kind of knowledge base contains the actual values of the semantic categories for each discourse entity. There are two kinds of propositions in this knowledge base. IsA(entity, type) defines the type of the entity, which is the same as the Name in the "lexicon". Each Prop(FeatureName, entity, value) defines a value for a semantic feature. Here is an example of items in this kind of knowledge:

IsA(HOME1_KITCHEN1, room)
Prop(Size, huge)
Prop(Light, bright)
Prop(Style, modern)

Given these three kinds of knowledge bases, our contrastive analysis algorithm can be summarized as:

a. Given a discourse entity x, find in the history list discourse entities of the same kind, by matching the part of speech, number, and gender defined by knowledge base 1 and put those entities in a list L[4].
b. Find in the knowledge base 3 alternative entities for x and put those entities in list L. The method of finding alternative entities is similar to that in (a), except that it also checks the Location value of the entity – only consider those entities that is in the same Location as entity x.
c. For each entity y in L, compare the semantic values defined by knowledge base 3 to those of the current entity. If mismatch is found on semantic category S, form a proposition contrastive(x, y, S).
d. Only keep those propositions that contain the entity that is more recent than others, since we assume contrast only happens in local context.

[4] We assume that contrast only happens within the same type of discourse entities.

The pragmatics produced by this contrastive analysis procedure can be used not only to determine the contrastive stress in the speech utterance, but also to indicate the need for generating gestures. Moreover, a similar procedure can analyze the surprising semantic values of a given discourse entity by comparing the actual semantic values defined by knowledge base 3 to sets of common values for the same semantic category defined by knowledge base 2. The result can be represented by a pragmatics in the form of “surprising(Object, Feature)”. This pragmatics is used to determine the generation of complementary gestures, as described in the previous chapter.

5. Implementation of Generation in REA

Although different speech and gesture generation systems may follow the same principles, the generation mechanism may vary depending on different system structure and how the system represents knowledge and dynamic context. This chapter describes the design and implementation of the paired speech and gesture generation framework in a particular embodied conversational system -- REA. First, I introduce the REA project and the system architecture. Detailed description of the generation framework follows. Finally, I present two examples that walk through the generation procedure.

5.1 REA – An Introduction

The domain of our embodied conversational agent is real estate. The agent, a 3D animated life-size female character displayed on a large projection screen, is able to talk to the user about real-estate properties, show those properties on the screen based on the user's request, and handle both interactional and propositional information exchange during the conversation through multiple communicative modalities.

The agent, named REA (which stands for real-estate agent) can perceive the user's behaviors through two video cameras that are mounted on the screen and two microphones that the user wears. A computer vision component constantly analyzes the stereo video input from the video cameras and converts them into meaningful non-verbal inputs such as the presence of the user, the location of the user, the gesturing status of the user, etc. A speech recognition component based on context-free grammar converts the user's speech into text and corresponding speech acts, while a separate process constantly measures the environment noise level and the tries to detect when the user starts speaking or pauses. The user input is interpreted as discourse functions and then converted into obligations that the agent needs to fulfill. The generation framework then generates appropriate speech and gestures. The agent's communicative behaviors are finally scheduled and carried out by a 3D animator component and a speech synthesis component. Below is a except from a sample interaction between the REA agent and the user:

> Tim approaches the projection screen. Rea is currently turned side on and is gazing idly about. As Tim moves within range of the two cameras mounted above the screen, Rea turns to face him and Tim started the conversation:

TIM> Hello, Rea.
REA> Hello. How can I help you?
TIM> I'm looking to buy a place near MIT.
Rea nods, indicating that she is following.
REA> I have a house to show you.
A picture of a house appears on-screen behind Rea who blinks and looks at the house and then at Tim.
REA> It's in Somerville.
TIM> Tell me about it.
Rea looks up and away while she plans what to say.
REA> It's bright and sunny.
Rea makes an expansive gesture with her hands.
Tim brings his hands up as if he is about to speak and so Rea does not continue; instead waiting for him to speak.
TIM> Tell me more about it
REA> It has a nice garden.
Rea sketches a curved gesture with her hands indicating that the garden extends along two sides of the house.
TIM> How far is it?
REA> It is five minutes to the Porter Square T station.
Rea makes it clear that it is five minutes ***on foot*** *from the T station by making a walking gesture.*
TIM> How big is the house?
REA> It has four bedrooms, three bathrooms. . .
Tim speaks to interrupt Rea who stops speaking immediately.
TIM> Wait. Tell me, Where is the master bedroom?
REA> I'm sorry, I didn't catch that. What did you ask me?
TIM> Where is the master bedroom?
REA> It's upstairs.
Rea points up
TIM> Where is the master bathroom?
REA> It's next to the bedroom.
Rea brings her hands together to indicate the relationship between the bedroom and the bathroom.
And the house tour continues.

5.2 System Architecture

To meet the requirement of handling interactional cues and understanding and generation of propositional content in real-time, we developed a modular system architecture for REA, as shown in figure 5.1. The architecture follows sequential processing of user input (Cassell et al., 1999). Each module in the architecture deals with different processing stage of the discourse, while all the modules share the same discourse model and communicate with each other using the same protocol. Using this architecture mimics the symmetrical processing (both input and output) of verbal and non-verbal behaviors in human face-to-face communication, and also ensures the scalability of the system from an engineering point of view.

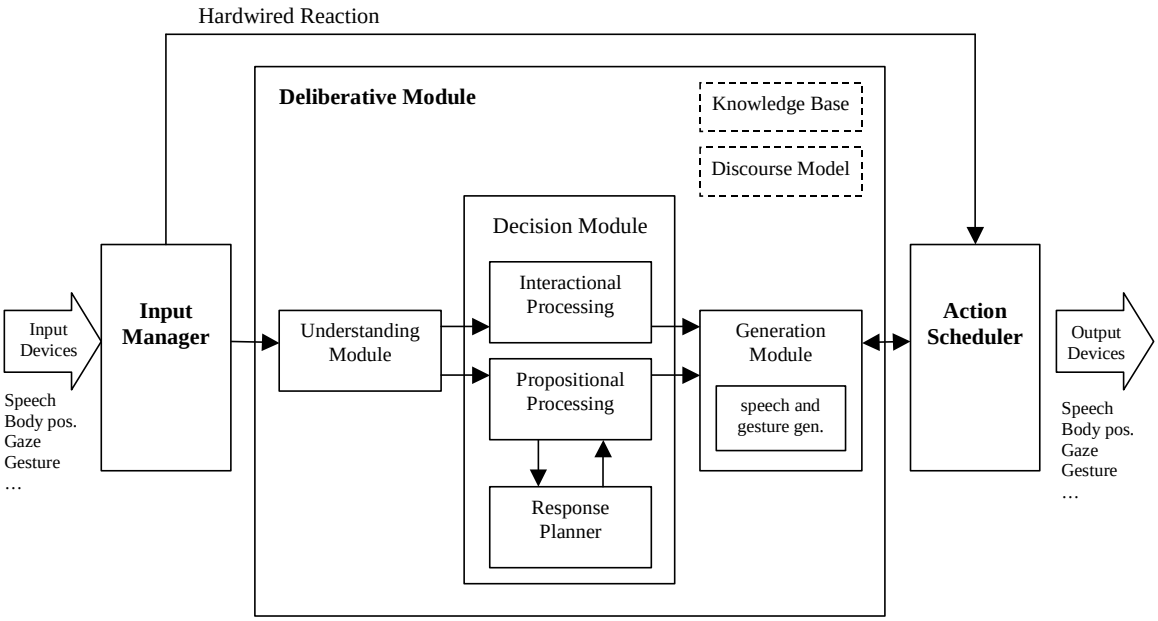


Figure 5.1: REA System Architecture

The Input Manager handles the incoming data stream from various input devices. The input data stream currently includes the video stream from the stereo vision, the speech input from the speech recognition program, and the audio threshold processing program. The Input Manager converts these data into forms that other modules in the system can handle (such as speech text, user location, user gesture, etc.) and passes them to other modules. Some of the input data can be directly routed to the Action Scheduler through the Hardwired Reaction, which enables the agent to respond immediately to certain user behaviors that require fast reaction. For example, the user location information is immediately handled to allow the agent to move her eyes following the user.

The deliberative processing of the system is accomplished by the Understanding Module, Decision Module, and Generation Module, with reference to a global knowledge base and a discourse model. These processing units track the user's focus, generate appropriate responses, and ensure the coherence in the conversation. The Understanding Module assembles user input data from all modalities that have been passed from the Input Manager and translates them into interactional and propositional discourse functions, which reflect the user's intention. These discourse functions are then routed to the Decision Module. The interactional processing sub-module inside the Decision Module is responsible for updating the conversational state, i.e. who

started the conversation, who has the turn, whether the conversation has been put on hold, etc. Figure 5.2 shows the shifting of conversational states.

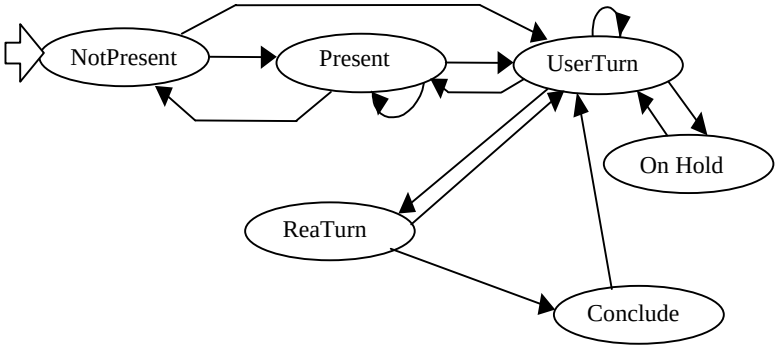


Figure 5.2: Conversational States in REA

The Propositional Processing sub-module and the Response Planner sub-module inside the Decision Module are responsible for updating the content exchange in the conversation. They keep track of the user's propositional input (in the form of Speech Acts), update the context (such as the information structure, attentional states, and other pragmatics), maintain the common ground (i.e. shared knowledge) between the user and the agent, and formulate series of actions (in the form of communicative goals). The Generation Module accepts the communicative goals and context updates from the Decision Module and realize the communicative goals into a description of communicative actions which include speech, gesture, or a combination of both.

The Action Scheduler coordinates actions descriptions output from the Generation Module at the lowest level. It takes a set of modality-specific commands and executes them through 3D animator and speech synthesis components in a synchronized way. This is realized through the use of event conditions, which define when the action should be executed.

The system is currently implemented in C++ and CLIPS and is largely rule-based. The modularity of the system is made possible by using the KQML communication protocol that is specialized for inter-agent communication (Finin and Fritzon, 1994). Below is a sample KQML message in the REA system:

```
(tell :sender UM :recipient DM :content
  (commact :sender USER :recipient REA
    :input [(speaking :state TRUE)
            (gesturing :state TRUE) ]
```

```
    :prop NONE
    :intr [ (takingturn) ]
  )
)
```

Figure 5.3. A Sample KQML Performative

5.3 The Generation Framework

As required by the fact that human face-to-face conversation is a bi-directional process, the REA system is a symmetrical system, i.e. it handles multiple input modalities as well as generating behaviors in multiple modalities. The internal processing of the system is function-based and modality independent. It is the generation module that realizes internal discourse functions (or, communicative goals) into surface behaviors in appropriate modalities, such as speech and gesture, and in appropriate forms. These generated behaviors need to convey domain propositions that encode specified kinds of information about a specified object or event, and also fit the context specified by the system, to the best extent possible (Cassell & Stone, 1999). As shown in figure 5.4, the generation process is centered around the SPUD (Stone& Doran, 1997) natural language generator, which is capable of building a sentence step by step, adding a lexical item at each step that meets the communicative goals, and the context constraint specified by the discourse manager, and the syntactic and semantic constraint provided by the contextual background.

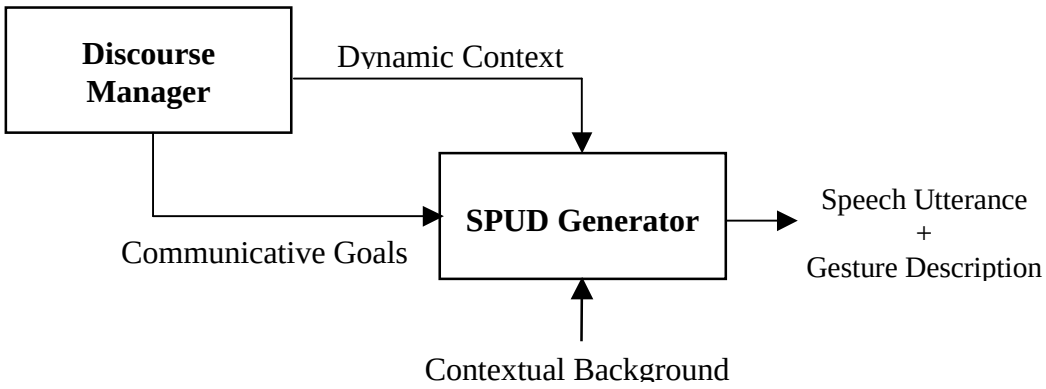


Figure 5.4: Generation Module Overview

Instead of choosing an entire the gesture from some sort of a gesture dictionary, the REA generation framework tries to synthesize a gesture from basic elements of gesture, such as number of hands, hand shapes, trajectories, and starting positions for both hands. Most of these elements are linked to semantic features specified by the contextual background. In this way, we

avoid the pain of building a large gesture dictionary. Instead, we just need to specify limited number of lexical items for basic gesture elements. Moreover, the system treats the lexical items in speech and descriptions of gesture elements equally, as “lexical descriptors”. Thus, the speech and gesture can be derived from the same underlying knowledge representation.

Figure 5.5 shows the detailed architecture of the paired speech and gesture generation framework in the REA system. The generation process starts when the system’s decision module sends out a generation speech act. The generation speech act is usually in “describe(object, aspect)” form. This speech act is then converted by the Request Formulator into a SPUD communicative goal, “describe s(f_object_aspect)”. The communicative goal basically tells the SPUD generator to construct an utterance that describes the f_object_aspect fact based on its static contextual background and dynamic discourse context.

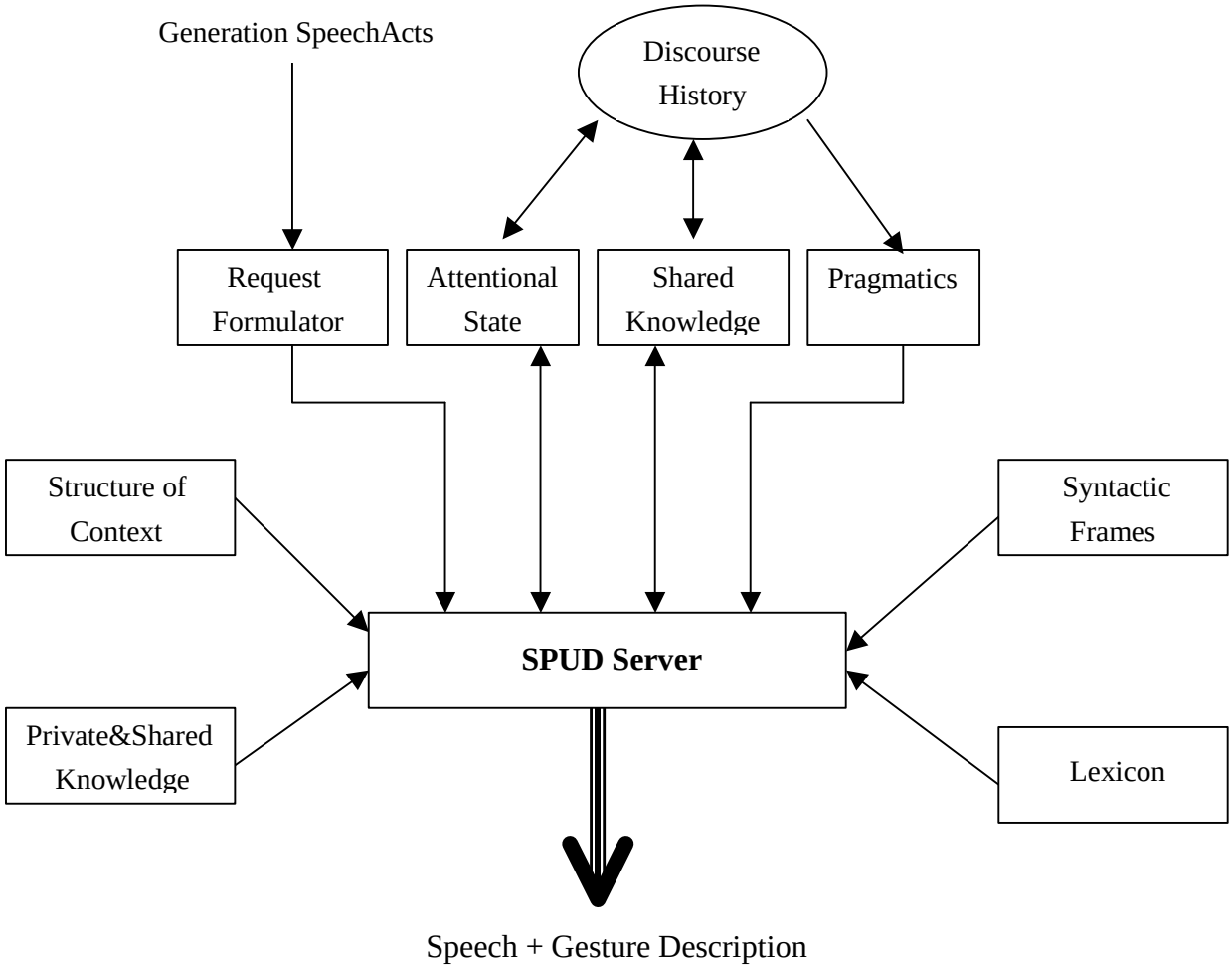


Figure 5.5: REA Generation Architecture

The structure of context, Private and Shared Knowledge, Syntactic Frames, and Lexicon form the static contextual background from which the SPUD generator can draw for its communicative content. The Structure of Context and Private and Shared Knowledge module contain the system's knowledge about discourse entities in the form of fact propositions and the organization of those propositions. The propositions in the system are divided into groups. They are classified into categories such as Type, Impression, Config, etc. In addition, the Structure of Context sets up an inheritance relationship among those categories. For example, propositions in the Shared category define knowledge that the user and the system share, other categories contain all knowledge from the Shared categories and other specific knowledge propositions, and all these propositions are contained in the systems Private type of knowledge. This organization of knowledge defines the common background.

The Syntactic Frames and Lexicon define the grammar for the generation system. The SPUD syntax is a hierarchical structure, formalized using Feature-Based Lexicalized Tree Adjoining Grammar (LTAG) (Stone & Doran, 1997). The Syntactic Frames is a grammar file that contains syntactic structure for possible utterances. The syntactic structure is in a hierarchical tree format. Each entry in the tree structure contains at least one lexical item. The gesture and intonation are realized in sub-trees using the substitution operation in the LTAG. In the substitution operation, a leaf node in the tree structure is replaced by the root node of a sub-tree. Moreover, these syntactic structures are linked with communicative goals and pragmatics checks. The lexicon provides the actual lexical item for the generation. In the system, words in speech, components of gesture description, and intonation description are defined in the same way. They are all lexical items in the lexicon. Each lexical item is specified in three parts. The first part is the syntax of the element, which lays out the contributions the lexical item would make to an utterance that contains it. The second part is the semantics, which defines the content that the lexical item carries. It sets up a link to the knowledge base because it is specified in the same proposition form as the entries in the knowledge base. The third part is the pragmatics conditions, which is a context check before the lexical item can be used in a lexical entry in the syntax tree structure.

During the conversation, the SPUD generator gets dynamic updates from the systems discourse model to stay on top of the changing state and context of conversation. These updates include the current attentional state, shared knowledge updates to the common ground, and pragmatics. The attentional state updates the attentional prominence for each discourse entity, i.e. what entities are as salient or as likely to be referenced as the discourse entity. The system's history list contains

all discourse entities previously mentioned in the discourse ranked by attentional prominence. After each generation process, the system assumes that the information carried by the utterance has become a belief of the user. Therefore, all the propositions used in the generation process are put into the Shared category, so that the generation module does not use them again as new information. Pragmatics plays an important role in the whole generation process. They represent most of the context and constrain the use of syntax entries. During the conversation, the system keeps analyzing and updating the following pragmatic information:

a. Information structure, in the form of theme(object) and rheme(object). The mechanism of analyzing information structure is discussed in section 4.1. These pragmatics constrain both gesture and intonation generation.
b. Surprising semantic features, in the form of Surprising(object, feature). It is useful in determining the need to generate a complementary gesture. The mechanism of analyzing surprising semantic features is discussed in section 4.3.
c. Contrast, in the form of Contrastive(object1, object2, feature). It is useful for determining contrastive stress in intonation, as well as the need for generating contrast gestures. Section 4.3 discussed the contrastive analysis in the system.
d. New/given information, in the form of new(object) and activated(object). This information is helpful in the generation of pronouns. The new/given information can be obtained simply by looking up current object in focus in the discourse history list.
e. Mutually observability[5], in the form of mutuallyobservable(object). This pragmatics specifies whether an object is mutually observable by the agent and the user. The system keeps track of which room the agent is in and it looks up in a table to determine if the object in focus is in the current room. This information is especially useful for generating some deictic gestures.

Based on the communicative goals, contextual background, and the dynamic context, the SPUD generator builds the utterance element by element; at each stage of construction, SPUD incrementally and recursively applies lexical specifications to determine which entities to describe and what information to include about them. If the generation process is successful, a speech utterance tagged with gesture descriptions and intonation descriptions is generated. It is then placed in a KQML frame and sent to the Action Scheduler for execution.

[5] This pragmatics is not an implementation of results from the data analysis. In the current system interaction, the agent shows a 3D model of the house or particular room while she talks about it.

The realization of complementary or redundant gestures in the paired speech and gesture generation framework is determined together by pragmatics and rules for organizing knowledge and communicative goals. We discussed the pragmatic constraints; here are the rules for organizing knowledge and communicative goals:

a. Grouping rules, which determine semantic features of an entity or an action that can be articulated together. These rules are realized as the categorization of semantic propositions and a secondary communicative goal that is sent along with the main communicative goal at the beginning of the generation. The categorization of semantic features, i.e. putting semantic features into the same communication category, is determined according to the results from the house description experiment. For example, propositions about the existence and shape of an object can be placed into the Introduction category. The secondary communicative goal is in the form of "communicateAll X $", which tells the SPUD generator to apply as many propositions in the X category as possible.
b. Appropriateness rules, which determine which semantic features are appropriate or easier to express through the gesture channel, and if appropriate, which gesture can best represent the semantic feature. These rules come from the statistics of semantic features in house description gestures, especially from the percentages of semantic features shown in Figure 3.2. See detailed discussion in section 3.3.

If there is a fact (proposition) in the category chosen by the secondary communicative goal(s) other than the fact determined by the primary communicative goal, and the fact (or combination of facts) is suitable to be articulated by a gesture, then the description of the gesture is filled into the gesture description slot. Therefore, SPUD generates the speech about the primary fact along with a complementary gesture that conveys more information. If there is no more facts in the category specified in the secondary communicative goal other than the fact determined by the primary communicative goal, but the primary fact is suitable to be articulated by a gesture, then the description of the gesture is filled into the gesture description slot. In this case, SPUD generates the speech about the primary fact along with a redundant gesture. If there is no more facts in the category specified in the secondary communicative goal other than the fact determined by the primary communicative goal, and the primary fact is not suitable to be

articulated by a gesture, then no gesture that carries meaning is filled into the gesture description slot. In this case, a default beat is inserted.

5.4 Two Examples

This section walks through the generation steps of two particular utterances that further explain the mechanism of the generation framework. The two utterances discussed are for introducing an object and for contrast on two different styles respectively.

5.4.1 Example 1 – Introduction

In the first example, the user and REA are discussing a house displayed on the screen, behind REA. The user says: “Tell me about the living room”. This utterance is converted into an input speech act by the system’s Understanding Module:

SA-REQUEST-INFO house2_livingroom1 impression

It is further mapped into an obligation of the agent, which is describing the impression of the living room:

SA-DESCRIBE house2_livingroom1 impression

There are usually many ways to realize a communicative goal. For example, to describe the impression about a room entity, the system could introduce new objects related to the entity, describe lighting condition of the entity, describe style of the entity, etc. The system’s prepositional processing module contains tables that define the rules of mapping communicative goals into more specific communicative goals. By looking at those tables, the prepositional processing module chooses to introduce a new object house2_livingroom1_object1, which is a staircase:

SA-INTRODUCE house2_livingroom1_object1

This speech act is then further turned into two a primary and a secondary communicative goals for the SPUD generator by the request formulator:

describe s(f_house2_livingroom1_object1_introuduction)
communicateAll Introduction $

At the same time, the system updates the context:

context domain(house2_livingroom1_object1, house2_livingroom1) ;attentional state

context domain(house2_livingroom1_object1, house2_livingroom1_object1) ;attentional state
context domain(house2_livingroom1, house2_livingroom1) ;attentional state
context activated(house2_livingroom1) ;given/new
context new(house2_livingroom1_object1) ;given/new
context rheme(house2_livingroom1_object1) ;information structure
context theme(house2_livingroom1) ;information structure
context surprising(house2_livingroom1_object1, Shape) ;surprising semantic feature

In the SPUD generator, the following syntactic structure is chosen to construct the utterance, in which G is a sub-trees for realizing gestures. x is fitted by house2_livingroom1_object1, and p is set to house2_livingroom1.

This syntactic structure is associated with semantic and pragmatics constraints:

Semantics: in(x,p)

Pragmatics: rheme(x) & surprising(x,Shape) & given(p) & ~mutuallyObservable(x)

The current context provides all the necessary pragmatic information and the system knowledge about house2_livingroom1 and house2_livingroom1_object1 fulfills the semantic constraint. Also in the Introduction category of system knowledge, there is an extra proposition about house2_livingroom1_object1, which is therefore chosen to be expressed by gesture G:

Introduction type(f_home2_livingroom1_object1_shape, home2_livingroom1_object1, spiral)

This spiral shape of the staircase is also mapped to the trajectory of the gesture G.

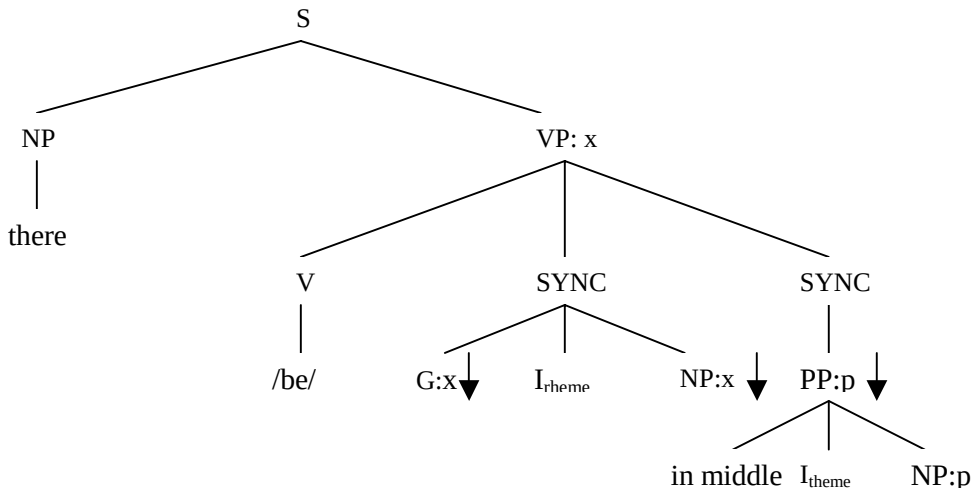


Figure 5.6: Syntactic structure 1

The I_{theme} and I_{rheme} specify theme and rheme accents respectively. Intonation specifications are realized as associated components in the syntax tree, because pitch accents only apply to words. Moreover, the pragmatics constraints for intonation are associated with the whole syntax tree. Thus, different intonation contours for the same speech are realized by specifying different trees with different pragmatic contraints.

Finally, after performing all substitution operations and realizing lexical items into surface form, the generation framework outputs a comprehensive description of communicative action that contains the speech text, intonation description, and the gesture description, as follows:

> SENTENCESTART there is <hands = one, right_handshape = pointer, left_handshape = none, right_trajectory = spiral_up, left_trajectory = none, left_startpoint = none, right_startpoint = center right center> a <accent, "H*"> staircase in the middle of <accent, "L+H*">it.

Figure 5.7 shows the agent performing the gesture, in which REA moves her right hand up spirally and keeps a pointing-up handshape.

Figure 5.7: "There is a staircase in the middle of it."

5.4.2 Example2 – Contrast

In the second example, the user and REA are discussing the same house. After asking REA to show the living room, the user asks: "Could you tell me more about the living room?" After the

processing by the input modules, this utterance is mapped into the same obligation of the agent as that in the previous example:

SA-DESCRIBE house2_livingroom1 impression

This time the prepositional processing module decides to introduce another new object house2_livingroom1_object2, which is hardwood floor:

SA-INTRODUCE house2_livingroom1_object2

Similarly, this speech act is then turned into two communicative goals for SPUD:

describe s(f_house2_livingroom1_object2_introuduction)
communicateAll Introduction $

The system then updates the context:

context domain(house2_livingroom1_object2, house2_livingroom1) ;attentional state
context domain(house2_livingroom1_object2, house2_livingroom1_object1) ;attentional state
context domain(house2_livingroom1_object2, house2_livingroom1_object2) ;attentional state
context domain(house2_livingroom1, house2_livingroom1) ;attentional state
context activated(house2_livingroom1) ;given/new
context new(house2_livingroom1_object2) ;given/new
context rheme(house2_livingroom1_object2) ;information structure
context theme(house2_livingroom1) ;information structure
context contrastive(house2_livingroom1_object2, house2_livingroom1_object3, Style)
;contrast

The system uses the algorithm discussed in section 4.3 to find out the contrast information. Following is a brief description of the actual process. In knowledge base 3, the entries about house2_livingroom1_object2 are:

IsA(HOME2_LIVINGROOM1_OBJECT2, room_amenity)
Prop(Style, victorian)
Prop(Location, HOME2_LIVINGROOM1_OBJECT2, HOME2_LIVINGROOM1)

The entry in knowledge base 1 related to the house2_livingroom1_object2 is:

(lexicon room_amenity NP single neutral Location Style)

In the discourse history list the system does not find an entity whose entry in knowledge base 1 matches the Name, Part of Speech, Number, and Gender field of house2_livingroom1_object2.

However, the system finds a matching object house2_livingroom1_object3 in knowledge base 3, which is wire for the Internet. The entry in the knowledge base 1 related to the house2_livingroom1_object3 is the same as that of house2_livingroom1_object2. The entries related to house2_livingroom1_object3 in knowledge base 3 are:

IsA(HOME2_LIVINGROOM1_OBJECT3, room_amenity)

Prop(Style, modern)

Prop(Location, HOME2_LIVINGROOM1_OBJECT3, HOME2_LIVINGROOM1)

Therefore, the house2_livingroom1_object3 becomes the only member of list L. Further comparison of other properties of these two objects found that they contrast on property Style. Thus a context proposition

contrastive(house2_livingroom1_object2, house2_livingroom1_object3, Style)

is produced.

In the SPUD generator, the following syntactic structure is chosen to construct the utterance. x is fitted by house2_livingroom1_object2, y is fitted by house2_livingroom1_object3 and o is set to house2_livingroom1. This syntactic structure is different from the one in the previous example in that it contains an auxiliary tree besides the S structure. The auxiliary tree is not necessary for the utterance generation and an utterance in the form of "o has x" is generated. However, in certain situation (in this case contrastive(x, y, Style)), the two trees can be combined together through the adjunction operation in the LTAG grammar. When the adjunction operation happens, an utterance in the form of "o has x but also y" is generated.

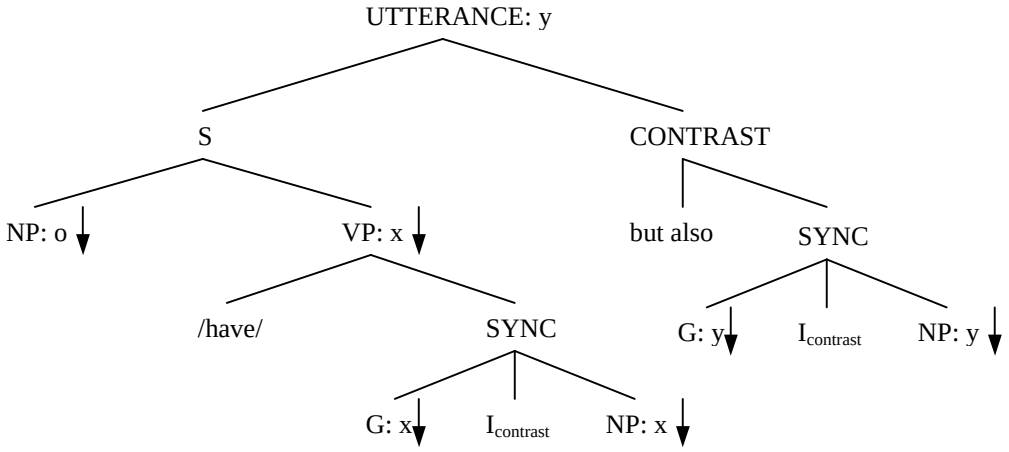


Figure 5.8: Syntactic structure 2

The semantic and pragmatics constraints associated with this syntactic structure are:

Semantics: have(o, x)

have(o, y)

Pragmatics: rheme(x) & given(o) & contrastive(x, y, Style)

The two gesture trees in the syntactic structure are associated with left hand beat and right hand beat respectively. The $I_{contrast}$ intonation is the same as I_{rheme}, which is a high pitch accent.

After performing all semantic checks, pragmatic checks, substitution and adjunction operations, the generation framework outputs the following comprehensive description of communicative action:

SENTENCESTART It has <hands = one, right_handshpe = beat, left_handshape = none,…> <accent "H*"> Victorian styles but also <hands = one, right_handshpe = none, left_handshape = beat,…> <accent "H*"> modern amentities.

Figure 5.9 shows the agent performing the two beats to show the contrast on styles:

Figure 5.9: "It has Victorian styles but also modern amenities"

5.5 Evaluation of Implementation

Approximately 20 utterances have been generated through the paired speech and gesture generation framework. Appendix B lists all the generation results[6]. The communicative goals of

[6] To facilitate reading, the gesture descriptions are translated into natural language from raw specifications of gesture components.

the utterances range from introducing an object, describing location of an object, to describing general impression about an object. The description of gestures is realized into actual gesture by the 3D animation component of the REA system. The speech along with intonation specifications are passed to and produced by a custom-built version of Festival Speech synthesis system. Appendix C shows a few examples of REA's gestures. The generation framework is capable of generating utterances in near real-time – approximately 1.5 seconds for an utterance. However, the initialization of the SPUD knowledge bases takes a long time. Thus, if the topic of the conversation switches to a different real-estate property and new SPUD knowledge bases need to be reloaded, the system displays an obvious long delay.

The hierarchical mechanism of gesture generation works well. It generates appropriate representational gestures. The type of gesture and the complementary or redundant information expressed by gesture are determined together by the appropriateness of expressing semantic features in the gesture modalities, primary and secondary communicative goals, and other pragmatic information such as surprising semantic features and contrast. These rules come from the correlational study discussed in chapter 3. We tested our rules in the data, especially heuristics that use communicative goals to predict complementary or redundant information in gestures. The prediction accuracy of the location heuristic is 80.0%, while the two introduction heuristics successfully predict 78.8% and 93.8% gestures respectively. Analysis of those gestures that those rules fail to predict suggests that there might be other factors in determining the gesture generation, such as the contribution to the overall goal of the conversation.

Previous studies, especially (Torres, 1997), also investigate similar rules. However, they have not considered the role of the communicative goal, so that they do not explain the situation to use those rules. For example, Torres claim that distinctions of meaning about the shape of an object are realized in gesture. However, doing a gesture along with the mentioning of an object is not always right. Only in certain situations, such as introducing an object, can we apply the shape rule. Communicative goal defines the situation and therefore determines the occurrence and type of gesture.

The synthesis method of generating a gesture gives us great flexibility and scalability. All gestures except beats share the same gesture description sub-tree. Since there is no need to specify hand shape or trajectory for beats, there is a simpler sub-tree for them. We assume that a gesture is composed of some basic form components and those components are associated with

semantics. However, our specification of the association between semantics and gesture components is ad hoc – it is just based on our individual observation from the experiment data and there is no general rule that guides the specification process. Usually, a semantic feature can determine a major form component of a gesture, such as hand shape or trajectory. The relation between semantics and the rest of the gesture, such as starting position, is not apparent. Further study is needed to research gesture morphology and its determining factors.

The generated f_0 specifications for pitch accent produce intonation-rich speech through the Festival speech synthesis system. It is easy to tell the difference between an utterance with and without intonation specifications. There is one case in which contrastive stress is generated. When talking about the general impression about the bedroom of a house, the system generates "It has Victorian style but also modern amenities." A contrast is found between "Victorian" and "Modern", therefore pitch accents are put on these words besides the noun discourse entities "style" and "amenities". However, while the contrast analysis algorithm described in chapter 4 works, we find that specifying alternative set for each semantic feature of each discourse entity is not a trivial task. When the system gets larger, those alternative sets may get intractable. In addition, we find that the current generation framework lack the mechanism to predict the duration change in speech utterances. More pragmatic information may be explored in the generation framework. For example, people often elongate a particular word in an utterance for emphasis or expression of an emotion. It is also common to stretch the duration of speech or add pauses when producing a time-consuming gesture.

6. Conclusion

Using face-to-face conversation as an interface metaphor, embodied conversational agents are likely to be easier to use than traditional graphical user interfaces. To make a believable agent that to some extent has the same social and conversational skills as humans do, the embodied conversational agent system must be able to deal with user's inputs from different communication modalities such as speech and gesture, as well as generating appropriate behaviors for those communication modalities. Moreover, research has shown that gesture and speech are integral parts of the whole communicative action in face-to-face conversation. Therefore, we should attempt to pair the generation of speech and generation of gesture in an embodied conversational agent system.

The thesis work is the first attempt to build a paired speech and gesture generation framework that can generate comprehensive descriptions of communicative actions in the real estate domain, including appropriate speech, gesture, and intonation. This framework, based on the SPUD natural language generation system, treats equally lexical items in speech, components in gesture, and components in intonation. They are represented in the same form and are described and constrained by the same static knowledge base and dynamic context. Specifically the discourse information structure and other pragmatic information determine where gesture and intonation occurs, while the communicative goal, the organization of knowledge, and domain constraints decide what semantic meaning gesture carries.

A correlational study of speech and gesture was conducted to investigate the distribution of semantics into speech and gesture modalities. I looked at this problem from a communication point of view and looked for the connection between the communicative goal and the semantic distribution across speech and gesture modalities. The results suggest that people use complementary gestures to manifest some uncommon features of the object in focus, to identify the object in focus from its alternatives, or to express other semantic information that is important for realizing the domain-specific overall communicative goal. Most redundant gestures are lexicalized and used to indicate the important semantic information that is key to realizing the overall communicative goal. Redundant gestures might serve to enhance the robustness of the communication.

The intonation generation in the paired speech and gesture generation framework is based on a combination of the Previous Mention Strategy and Information Structure Heuristics. The system is capable of generating f_0 specifications for pitch accents. Moreover, a contrastive analysis algorithm has been implementation to predict the contrast in semantic features and the contrastive stress in intonation. The generated intonation is produced through the Festival speech synthesis system. It is easy to tell the difference between an utterance produced with and without the intonation specifications.

The generation framework has been implemented in the REA system. The system operates on discourse functions. The generation module in the system realizes the output discourse functions (in speech act form) as speech and gesture behaviors. Simple discourse model has been built into the system to record discourse history, analyze and update information structure, surprising semantic features, and contrast, etc. The system currently generates about 20 different utterances in real-time. The generation framework can be easily scaled up to generate more gestures, with the addition of more semantic propositions, lexical items, and categorization rules.

7. Future Work

There is of course much future work to be done to improve the system. First of all, a larger scale and more thorough correlational study on speech and gesture needs to be conducted. Although I have analyzed more than 350 utterances, some results of the study are still not statistically significant. For example, in the study, I observed a few rules about semantics distribution in action description utterances. However, since the occurrence of total action description utterances in the house description domain is not enough, I cannot judge whether these rules are valid. A larger scale study across domains may provide more significant results. Moreover, although the communicative goal is crucial in determining the semantics distribution, it may not be the sole factor. Other factors such as word selection and intonation should be looked at in the larger scale study.

Secondly, in our generation system, we synthesize each gesture from several basic gesture components and associate those gesture components with semantics. This mechanism has been proven to be efficient in practice and convenient from engineering point of view. However, we have little evidence about whether this is a reasonable method and if the semantic relation to the gesture component can be generalized. More study on gesture morphology and semantics would be able to answer these questions.

The house description experiment shows that most of the utterances the subjects produce describe spatial relationships of rooms or objects in rooms. To be able to describe these spatial relationships, a full geometric representation of the scene is necessary. Based on the point of view of the agent, the system should be able to choose appropriate expressions to describe the location of an object.

Lastly, face-to-face conversation happens in real-time. This gives the embodied conversational agent time complexity requirement – all operations, including understanding of user's behavior, planning, generation, animation and speech production should happen in no more than 1 second. Currently, the processing of pragmatics requires a lot of computing cycles, and the SPUD generation takes even more time. The result is that an utterance takes about 1.5 seconds to be produced. The loading of databases during topic switches is even intolerable. How to reduce the computational complexity is a challenging research problem. Possible solutions include

optimizing the representation and organization of knowledge, and simplification of the natural language generation process.

Appendix A: Sample of Transcription and Analysis of the Experimental Data

Num	Time	N/L	SpLLch UttLranc L	Gestures	Communica tive Goals	Semantics in Speech & Gesture	Semantics in Speech	Semantics in Gesture	C/R	I/M/D	Memo
1	29:32	N	it's in a very nice residential neighborhood, it's really quiet. It's part of ... alongside of the street. So it's nice safe place.								
2	29:39	N	It get a nice little porch out in front, a porch kind of winds around whole one side, and part of the back and front of it.	right hand pointing up and circling half a circle	Shape	Shape	Shape	Shape	R	I	The gesture occurs at the second utterance, after the
3	29:49	N	um...(it) has trash cans out there, I think there was a hammock out there, which no one is using it because it was winter.								
4	29:55	N	um... you go through the front door, there is a nice stair case going up with carpet ...(can't tell)	right hand palm facing front and moving up	Introduction	Existance, Location, Direction	Existance, Location, Direction	Direction	R	I	up
5	30:00	N	um.. Off to your right is, I think that's the living room or kind of the family member area.	right hand move to right.	Introduction	Existance, Location	Existance, Location	Location	R	I	
6	30:07	N	um...it had, painted yellow, kind of not my favorite color, but, it had a fireplace in there, some lamps, you know, typical								
7	30:19	N	um.. Off to..if... instead of... If you going to the left after you came in the front door instead of the right,	two gestures, first two hands handshape S, circling in small circles, showing "going"	Action	Action	Action	Action	R	I	
8	30:19	N	um.. Off to..if... instead of... If you going to the left after you came in the front door instead of the	two gestures, second, right hand move and point to the left	Location	Direction	Direction	Direction	R	I	
9	30:26	N	Here, um.., wan't a stake (?), it was, actually I can't remember what this was in the video. It was um, oh, this was a library.that's right, a library that had a bunch of books there, had a nice, nice fireplace, just sit and read.								
10	30:43	S	um.. From the library, if you head off to the right, there was a nice dining room area there. Nice table, nice china cabinet of..a lot of nice dishes.	right hand move to right.	Introduction	Existance, Location	Existance, Location	Existance, Location	R	I	
11	30:52	N	it had a little mat out there,	right hand lifted to the center gesture space with palm facing front, wipe from left and right, right to left	Introduction	Existance, Location	Existance, Location	Location	C	I	mat is on a wall? Speech and gesture has same semantic feature, but not
12	30:53	N	look like at one point there is a fireplace, but it was boarded out. Um..., a bunch of female trinkets sitting on top of that.								

N/L: speakers – N: narrator L: listener
C/R: gesture type 1 -- complementary or redundant.
I/M/D: gesture type 2 – iconic, metaphoric, or deictic.

Num	Time	N/L	Speech Utterance	Gestures	Communicative Goals	Semantics in Speech & Gesture	Semantics in Speech	Semantics in Gesture	C/R	I/M/D	Memo
13	31:01	N	Um.. If you continue on through.., it was a kitchen.	left hand moving to the left	Introduction	Existance, Location	Existance	Location	C	I	the kitchen is on the left of the dining room. Only in gesture you know that.
15	31:22	N	Also (look) like (a) wine drinker, there is a wine rack and nice liquor cabinet and everything.	right hand palm facing from wiping at a lower level on wine rack and wiping at a higher level at a higher level	Introduction	Existance, Relative Position	Existance	Relative Position	C	I	the liquor cabinet is above the wine rack. And the intrinsic features in wiping trajectory, uprightness
16	31:27	N	um…then there is a staircase down to the basement.	left hand moving to left lower front	Introduction	Existance, Relative Position	Existance, Relative Position	Relative Position	R	I	??could be complementary on the relative position
17	31:31	N	and, like, a little door way down to the front porch.	right hand palm down, lifted a little and then down and move front	Configuration, Direction	Configuration,Direction	Configuration, Direction	Configuration,Direction	R	I	
18	31:33	N	it had a couple of plants hanging there. So there is lot of sunlight at the								
19	31:37	N	there were some large windows there.	two hands palm facing front, wiping	Introduction	Existance, Size, Location, Shape	Existance, Size, Location	Shape	C	I	Lexicalized "window"
20	31:39	N	underneath the staircase, zoom in, come right in, after the staircase, there is a bath, full bath.	two gestures, first one, right hand pointer handshape,	Location	Location, Action, Path	Location,	Action, Path	C/R	I	underneath.
21	31:39	N	underneath the staircase, zoom in, come right in, after the staircase, there is a bath, full bath.	two gestures, second, right hand demostrating "zoom in"	Action	Action	Action	Action	R	I	
22	31:48	N	um…nice things for like rose painted on it.								
23	31:51	N	um…once you went upstairs,	right hand pointer handshape, moving up	Action	Action, Direction, Path	Action, Direction, Path	Path	R	I	
24	31:55	N	um…if you went off to the right, there is a master bedroom there.	right hand move a little to the right, and beat	Action	Action, Direction	Action, Direction	Direction	R	I	
25	31:59	N	it got, nice, green, you know, very nice place of flower curtains.	right hand palm down, wave and move from center center center to peripher center	Introduction	Existance, Shape	Existance	Shape	C	I	First mention. Intrinsic feature of curtains "wrinkle"?
26	32:04	N	some kind of picture, I am not sure, it's kind of an oval shape thing	two hand demoing a circle, then two hands stretch out and the bottom	Shape	Shape	Shape	Shape	R	I	
27	32:16	N	If you, let's say if you pass the mster bedroom, if you went off to the right, there is a dressing room there.	two gestures, first, right hand palm facing left, moving forward	Action	Action	Action	Action	R	I	
28	32:16	N	If you, let's say if you pass the mster bedroom, if you went off to the right, there is a dressing room there.	two gestures, second, right hand move to the right	Action	Action, Direction	Action, Direction	Direction	R	I	
29	32:20	N	could also be another bedroom, they set it up as a dressing room, um, a bunch of clothes, everywhere. nothing much more cool about								

Speech utterances are transcribed word by word.
Some of them may be ungrammatical.

Appendix B: List of Generation Results

1. Speech: <accent L+H*> I have a <accent H*> condo.

   Gesture: Beat on the rheme – condo.

2. Speech: <accent L+H*> It is in a <accent H*> building in Boston.

   Gesture: Right hand, hand shape tapered O, moves up to periphery upper right. The gesture identifies the Lcoation feature – high up in the building.

3. Speech: <accent L+H*> It is <accent H*> bright and <accent H*> sunny.

   Gesture: Both hands move out, making an expansive gesture. The gesture indicates an important piece of information in the real estate domain – lighting.

4. Speech: <accent L+H*> It has <accent H*> two bedrooms, <accent H*> two bathrooms, and <accent H*> one study.

   Gesture: Beats on the rheme – numbers.

5. Speech: The <accent L+H*> condo is <accent H*> spacious.

   Gesture: Both hands move out, making an expansive gesture. The gesture indicates an important piece of information in the real estate domain – size.

6. Speech: <accent L+H*> It is just <accent H*> over the bridge from MIT.

   Gesture: Right hand, hand shape 4, palm facing down, moving forward along an arc trajectory. The gesture indicates an important piece of information in the real estate domain – location.

7. Speech: <accent L+H*> It is a <accent H*> modern kitchen.

   Gesture: Beat on the rheme – modern.

8. Speech: The <accent L+H*> study is <accent H*> wired for the Internet.

   Gesture: Beat on the rheme – wired for the Internet.

9. Speech: <accent L+H*> I have a <accent H*> house.

Gesture: Beat on the rheme – house. This utterance is generated based on exactly the same syntactic tree, information structure, and other pragmatic constraints as utterance 1.

10. Speech: <accent L+H*> It is in <accent H*> Somerville.

    Gesture: Right hand, hand shape G, palm facing down, move forward. The gesture shows complementary information about Somerville – it is far away.

11. Speech: The <accent L+H*> house has <accent H*> two stories and there is a <accent H*> staircase in the middle.

    Gesture: Right hand, pointing up, moving up along a spiral trajectory. The gesture shows the shape of the staircase. It is to identify the particular staircase from other types of staircases.

12. Speech: There is a <accent H*> chimney in the kitchen.

    Gesture: Right hand, cup hand shape, makes contact with left hand, which is B spread hand shape with palm facing right. The gesture shows the shape of the chimney and the relative of the chimney to the wall.

13. Speech: <accent L+H*> It has <accent H*> three bedrooms, <accent H*> two bathrooms, and <accent H*> one living room.

    Gesture: Beats on the rheme – numbers. This utterance is generated based on exactly the same syntactic tree, information structure, and other pragmatic constraints as utterance 4.

14. Speech: There is a <accent H*> chimney here.

    Gesture: Left hand pointing to the chimney in the picture showed on screen. Since the picture of the kitchen is shown on screen (so that both REA and the user can see the chimney), a deictic gesture is generated.

15. Speech: There is a <accent H*> Jacuzzi in the <accent L+H*> bathroom.

    Gesture: Both hands are lifted into the gesture space and demonstrate the “bulbbling water” feature of the Jacuzzi.

16. Speech: The <accent L+H*> bedroom is <accent H*> upstairs.

    Gesture: Right hand points up. The gesture is lexicalized with the word “upstairs”.

17. Speech: The <accent L+H*> house is <accent H*> five minutes to the Porter Square T station.

    Gesture: Right hand shows the walking information.

18. Speech: The <accent L+H*> bathroom is next to the bedroom.

    Gesture: Two hands face each other, with B spread hand shape. The gesture repeat the “next to” information in the speech. Spatial relation is important in the house description domain.

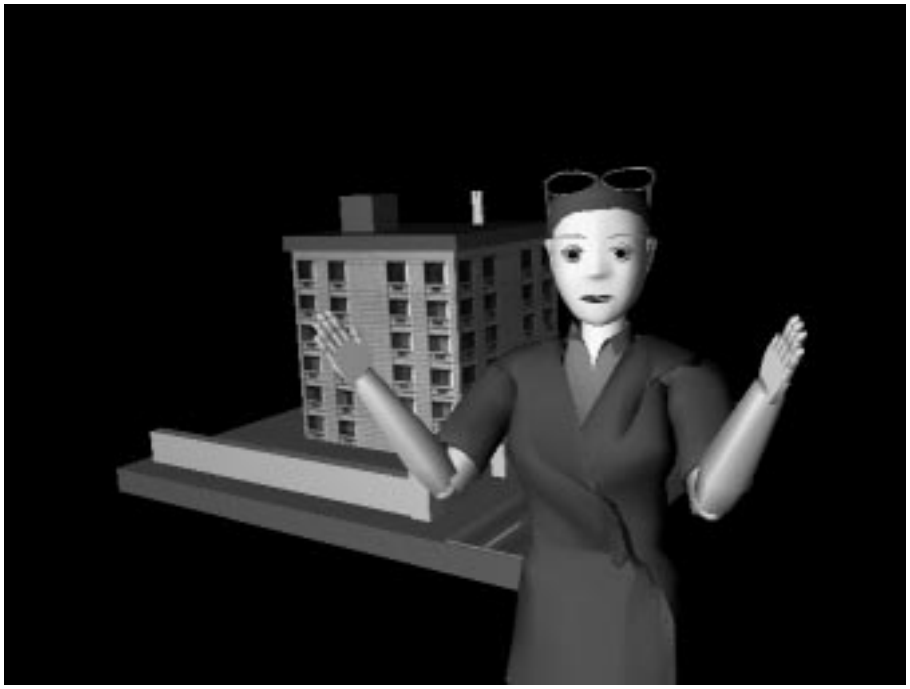
19. Speech: The <accent L+H*> living room has <accent H*> Victorian style but also <accent H*> modern amenities.

    Gesture: Right hand beats on the Victorian style and left hand beats on the modern amenities. The two beats shows the contrast on the “Victorian” and “modern”.

20. Speech: The <accent L+H*> living room has <accent H*> hardwood floors.

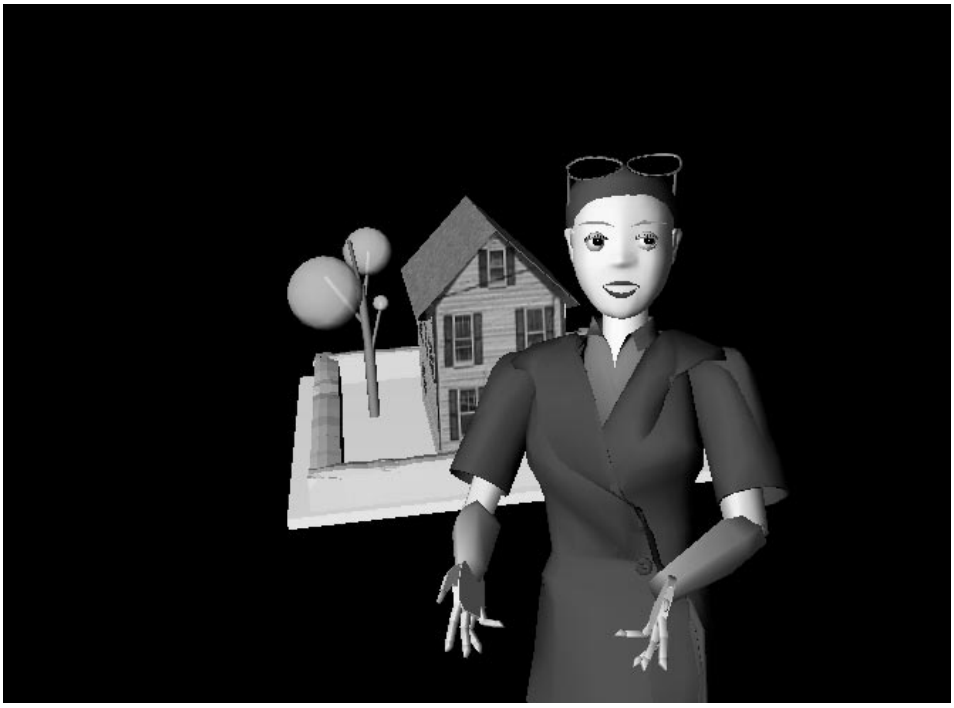
    Gesture: Beat on rheme – hardwood floors.

Appendix C: Sample of Generated Gestures

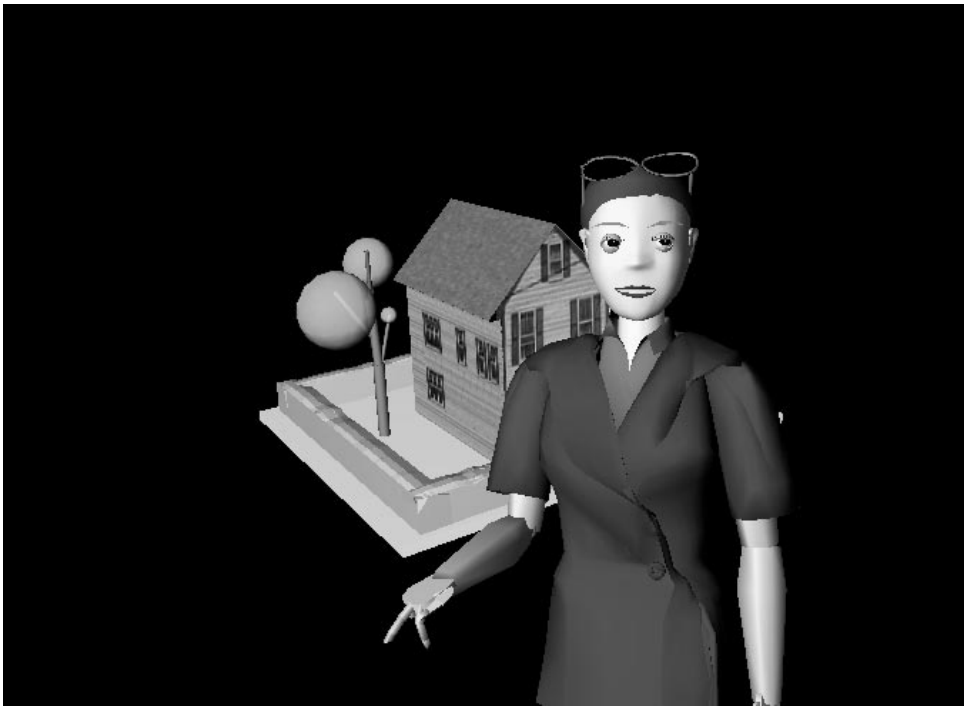


Speech: “It (the condo) is bright and sunny.”
Gesture: expansive gesture that expresses the impression of brightness.

Speech: “There is a chimney in it (the kitchen)”
Gesture: shows the shape of the chimney and its relative position to the wall.



Speech: “There is a jacuzzi in the bathroom.”
Gesture: lexicalized, demonstrates a typical feature of jacuzzi: bulbbing water.



Speech: “It (the house) is five minutes to the porter square T station”
Gesture: walking gesture adds that it is 5 minutes walking distance.

References

[Altenberg, 1987] Altenberg, B, Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion, Lund Studies in English, vol 76, Lund University Press, Lund.

[André, Rist, and Mueller, 1998] André E., Rist T., and Mueller, J., Integrating Reactive and Scripted Behaviors in a Life_like Presentation Agent, In Proceedings of Agents '98, Minneapolis/St. Paul, 1998.

[André and Rist, 2000] André E. and Rist T., Presenting Through Performing: On the Use of Multiple Lifelike Characters in Knowledge-Based Presentation Systems, in the proceedings of 2000 ACM Intelligent User Interfaces Conference, New Orleans, Louisiana, 2000

[Badler et al, 1999] Norman I. Badler, Rama Bindiganavale, Juliet Bourne, Jan Allbeck,Jianping Shi, and Martha Palmer, Real Time Virtual Humans, in the proceedings of the International Conference on Digital Media Futures, Bradford, UK, 1999

[Black and Taylor, 1997] Black, A. and Taylor, P., Festival Speech Synthesis System: system documentation (1.1.1) Human Communication Research Centre Technical Report HCRC/TR-83, 1997.

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

[Brown, 1983] Brown, G., Prosodic structure and the given/new distinction, in Ladd, D. R. and Cutler, A., editors, Prosody: Models and Measurements. Springer Verlag, Berlin, 1983.

[Butterworth and Hadar, 1989] Butterworth, B. and Hadar, U., Gesture, Speech, and Computational Stages: A Reply to McNeill, Psychological Review, 96:168-174, 1989.

[Cassell et al., 1994] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., & Stone, M. (1994). Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. Proceedings of Siggraph '94, Orlando.

[Cassell et al., 1999] Cassell, J., Bickmore, T, Campbell, L., Vilhjalmsson, H., and Yan, H., Conversation as a System Framework: Designing Embodied Conversational Agents, to appear in *Embodied Conversational Agents*, Cassell, J. eds, 1999, MIT Press.

[Cassell and Prevost, 1996] Cassell, J. and S. Prevost. "Distribution of Semantic Features Across Speech and Gesture by Humans and Computers." Proceedings of the Workshop on the Integration of Gesture in Language and Speech, 1996.

[Cassell, Torres, and Prevost, 1999] Cassell, J., Torres, O,. and Prevost, S. "Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation." In Wilks (ed.), Machine Conversations. The Hague: Kluwer, 1999.

[Cassell & Stone, 1999] Cassell, J. and Stone, M. "Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems." *AAAI 1999 Fall Symposium on Narrative Intelligence*.

[Chovil, 1992] Chovil, N. Discourse-Oriented Facial Displays in Conversation, in Research on Language and Social Interaction, Vol 25, PP163-194, 1992.

[Clark and Marshall, 1981] Clark, H. H., and Marshall, C. R., Definite Reference and Mutual Knowledge. In A. K. Joshi, B. L. Webber, & I. Sag (Editors), Elements of Discourse Understanding (pp. 10-63). Cambridge: Cambridge University Press, 1981.

[Feiner & McKeown, 1990] Feiner, S. andMcKeown, K.R., Coordinating text and graphics in explanation generation. In Proceedings of the AAAI-90, PP442-449, Boston, 1990

[Finin and Fritzon, 1994] Finin, T. and Fritzon, R., KQML as an agent communication language.in the Third International Conference on Information and Knowledge Management (CIKM, '94), Gaithersburg, Maryland, 1994

[Freedman, 1972] Freedman, N.,The Analysis of Movement Behavior During the Clinical Interview. In Siegman A. and Pope, B., editors, Studies in Dyadic Communication, Pergamon, New York, 1972.

[Goodwin, 1981] Goodwin, C., Conversational organization: interaction between speakers and hearers, 1981, New York: Academic Press.

[Green et al., 1998] Green, N. and Giuseppe C., Stephan K., Steven R., and Johanna M.. A Media-Independent Content Language for Integrated Text and Graphics Generation. Proceedings of the Workshop on Content Visualization and Intermedia Representations (CVIR'98) of the 17th International Conference on Computational Linguistics (COLING '98) and the 36th Annual Meeting of the Association for Computational Linguistics (ACL'98). Montreal, Canada, August 15, 1998.

[Grosz and Sidner, 1986] Grosz, B. and Sidner, C., Attention, Intentions, and the Structure of Discourse. Computational Linguistics, Vol. 12, No. 3, 1986.

[Hirschberg, 1990] Hirschberg, J., Accent and Discourse Context: Assigning Pitch Accent in Synthetic Speech. In proceedings of the Eighth National Conference on Artificial Intelligence, pages 952-957, 1990.

[Hiyakumoto et al., 1997] Hiyakumoto, L., Prevost, S., and Cassell, J., Semantic and Discourse Information for Text-to-Speech Intonation, in ACL Workshop on Concept-to-Speech Technology, 1997.

[Jackendoff, 1983] Jackendoff, R., *Semantics and Cognition*, The MIT Press, Cambridge, MA 1983.

[Kendon, 1972] Kendon, A., Some relationships between body motion and speech. In Siegman A. and Pope B. (eds.), Studies in dyadic communication, 177-210, 1972, Pergamon Press, New York.

[Kendon, 1990] Kendon, A., The negotiation of context in face-to-face interaction. In A. Duranti and C. Goodwin, editors, Rethinking context: Language as interactive phenomenon, 323–334, 1990 New York: Cambridge University Press.

[Kendon, 1994] Kendon, A., Do Gestures Communicate? A review. Research on Language and Social Interaction, 27(3): 175-200, 1994.

[Laver, 1975] Laver j. Communicative Functions of Phatic Communion, Kendon, A., Harris, R., and Key, M. editors, The Organization of Behavior in Face-to-Face Interaction, PP215-238, Mouton, The Hague, 1975.

[Lester et al., 1997] Lester J., Converse S., Stone B., Kahler S., and Barlow T., Animated Pedagogical Agents and Problem-Solving Effectiveness: A Large-Scale Empirical Evaluation, in the proceedings of the Eighth World Conference on Artificial Intelligence in Education, pp. 23-30, Kobe, Japan, 1997.

[Lester et al., 1999] Lester, J. C., Towns S. G., Callaway C. B., Voerman J. L., and FitzGerald P. J., Deictic and Emotive Communication in Animated Pedagogical Agents, to appear in *Embodied Conversational Agents,* Cassell, J. eds, 1999, MIT Press.

[McNeill, 1992] McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago.

[Monaghan, 1991] Monaghan, A., Intonation in a Text-to-Speech Conversion System. Ph.D. thesis, University of Edinburgh, 1991.

[Nobe et al., 1998]Nobe, S., Hayamizu, S, Hasegawa, O., and Takahashi H., Are Listeners Paying Attention to the Hand Gestures of an Anthropomorphic Agent? An Evaluation Using a Gaze Tracking Method, in Lecture notes in computer science, Vol 1371, 1998

[Pierrehumbert and Hirschberg, 1990] Pierrehumbert, J. and Hirschberg, J., The Meaning of Intonational Contours in the Interpretation of Discourse, in Cohen, P. Morgan, and Pollack, editors, Intentions in Communication, MIT Press, Cambridge MA, pp. 271-312, 1990.

[Perlin and Goldberg, 1996] K. Perlin and A. Goldberg. Improv: a system for interactive actors in virtual worlds. In Proceedings of SIGGRAPH 96, Computer Graphics Proceedings, Annual Conference Series, pages 205–216, 1996.

[Prevost and Steedman, 1994] Prevost, S. and Steedman, M., Specifying Intonation from Context for Speech Synthesis. Speech Communication, 15:139-153, 1994.

[Prevost, 1996] Prevost, S., Modeling Contrast in the Generation and Synthesis of Spoken Language, in ICSLP '96: the Fourth International Conference on Spoken Language Processing, Philadelphia, 1996

[Prince, 1981] Prince, E., Toward a taxonomy of given-new information, in Cole, P., editor, Radical Pragmatics, Academic Press, New York, 1981.

[Rickel and Johnson, 1999] Rickel J. and Johnson, W. L., Animated agents for procedural training in virtual reality: Perception, cognition and motor control. Applied Artificial Intelligence, 13:343–382, 1999.

[Rijpkema and Girard, 1991] Rijpkema, H. and Girard, M., Computer animation of hands and grasping. Computer Graphics, 25(4):339–348, 1991.

[Rogers, 1978] Rogers, W.T., The Contribution of Kinesic Illustrators towards the Comprehension of Verbal Behavior Within Utterances. Human Communication Research, 5, pages 54-62, 1978.

[Rosenfeld, 1987] Rosenfeld, H.M., Conversational Control of Nonverbal Behavior, in Siegman, A. and Feldstein, S. editors, Nonverbal Behavior and Communication, Lawrence Erlbaurn Associates, Inc, 1987.

[Steedman, 1999] Steedman, M., Information Structure and the Syntax-Phonology Interface, to appear in Linguistic Inquiry.

[Stone & Doran, 1997] Stone, M. and Doran, C., Sentence Planning as Description Using Tree-Adjoining Grammar. Proceedings of ACL 1997, pages 198--205.

[Torres, 1997] Torres, O.E. (1997). Producing Semantically Appropriate Gestures in Embodied Language Generation. MS thesis, Massachusetts Institute of Technology, Media Laboratory.

[Wahlster et al., 1991] Wahlster, W., André, E., Fraf, W., and Rist, T., Designing illustrated texts: How language production is influenced by graphics production. In Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics, PP 8-14, Berlin, 1991