USING EMOTION TO GAIN RAPPORT IN A SPOKEN DIALOG SYSTEM

JAIME CESAR ACOSTA

Department of Computer Science

APPROVED:

Nigel Ward, Chair, Ph.D.

David Novick, Ph.D.

Stephen Crites, Ph.D.

Patricia D. Witherspoon, Ph.D.
Dean of the Graduate School

to all of

my family

with love

USING EMOTION TO GAIN RAPPORT IN A SPOKEN DIALOG SYSTEM

by

JAIME CESAR ACOSTA, M.S.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

December 2009

Acknowledgments

I would like to thank God and my family for all of the support that was provided during this period in my life.

I would like to give my deepest appreciation to Nigel Ward for all of his help as an advisor. I thank David Novick and Stephen Crites for being on my committee and for their constructive feedback.

I am grateful to Anais Rivera for collecting the persuasive dialog corpus, Rafael Escalante-Ruiz for the persuasive letter generator system, and Jason Zheng for his help implementing the VoiceXML Prototype.

I would also like to thank Alex Vega, Shreyas Karkhedkar, and my wife, Tam Acosta for their help during testing and Ben Walker and Josh Mc Cartney for their help during the experiments.

This work was supported in part by an ACTEDS Scholarship, by NSF Awards IIS-0415150 and IIS-0914868, and by the US Army Research, Development and Engineering Command, via a subcontract to the USC Institute for Creative Technologies.

Abstract

Although spoken dialog systems are becoming more widespread, their application is today limited largely to domains involving simple information exchange. To enable future applications, such as persuasion, new capabilities are needed. One barrier to the creation of such applications has been the lack of methods for building rapport between spoken dialog systems and human users, and more generally the inability to model the emotional and interpersonal aspects of dialog. This dissertation focuses on improving this.

A corpus of persuasive dialogs that in which a graduate coordinator informed undergraduate students about the graduate school option was analyzed. Although much of each dialog was involved in conveying factual information, there was also a heavy use of what appear to be rapport-building strategies. This seemed to occur through emotional coloring of the utterances of both coordinator and students as heard in prosodic variation, including variation in pitch, timing, and volume.

Some of these rapport-building strategies were modeled and implemented in a spoken dialog system named Gracie (Graduate Coordinator with Immediate-Response Emotions). Gracie is the first dialog system that uses emotion in voice to build rapport with users. This is accomplished by first detecting emotions from the user's voice, not classic emotions such as sadness, anger, and joy, but the more subtle emotions that are more common in spontaneous conversations. These subtle emotions are described with a dimensional approach, using the three dimensions of activation (active/passive), evaluation (positive, negative), and power (dominant/submissive). Once the user's emotional state is recognized, Gracie chooses an appropriate emotional coloring for the response.

To test the value of such emotional responsiveness, an experiment with 36 subjects examined whether a spoken dialog system that recognizes human emotion and reacts with appropriate emotion can help gain rapport with humans. Users felt significantly more rapport with Gracie to the controls, and in addition, users significantly preferred Gracie

to the other two systems. This suggests that dialog systems that attempt to connect to users should vary their emotional coloring, as expressed through prosody, in response to the user's inferred emotional state.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Technologies such as automatic speech recognition and speech synthesis make it possible for spoken dialog systems to be used widely. However, spoken dialog systems are today largely limited to domains involving simple information exchange. This is largely due to the shortcomings of their underlying technologies. Speech recognition and speech synthesis are of course critical, but their performance is improving quickly and we are now approaching the point where other factors are limiting their further use, for example, in domains such as persuasion. One issue of interest is the inability of dialog systems to recognize and adapt to user emotional state, not only as revealed with words and topics, but also as expressed by nonverbal features such as prosodic variations.

1.1 Aims

This dissertation focuses, specifically, on extending dialog systems to enable them to gain rapport with users. This is accomplished by recognizing user emotional states and reacting with appropriate emotional coloring of responses.

1.1.1 Rapport

Gratch et al. [18] defines rapport as *a feeling of connectedness that seems to arise from rapid and contingent positive feedback between partners and is often associated with socio-emotional processes*. Previous work has investigated this phenomenon with virtual agents that react appropriately to human nonverbal behavior [9],[18],[8]. For example, Gratch et al. [18] has shown that automated agents are able to gain this “connectedness” with human users by nodding at the right moment. This dissertation looks at how rapport can be gained during voice-only interaction with a spoken dialog system.

Shepard et al.'s Communication Accommodation Theory (CAT) [36] serves as a basis for this work. The theory states that humans use prosody and other nonverbal behaviors in order to decrease social distance with an interlocutor (achieve convergence). Many have shown that prosody is an indicator for emotional state [11, 13, 22].

1.1.2 Emotional Intelligence

In order to gain rapport, beyond the simple responsive patterns investigated as mentioned above, one promising approach is to add "emotional intelligence." In this dissertation, the term emotional intelligence is defined as the ability of a spoken dialog system to detect emotion, determine an appropriate response, and to render the response to the user with appropriate emotional coloring. This emotional coloring refers to prosodic variations in voice.

1.2 Thesis Statement

The main hypothesis of this research is that a spoken dialog system with emotional intelligence is better at gaining rapport than a dialog system without emotional intelligence. To test the hypothesis, a spoken dialog system, Gracie (GRAduate Coordinator with Immediate-Response Emotions), was built and evaluated with users.

The rest of this dissertation is structured as follows. First a survey of related research is presented. Next, the steps that led to the main research question are described. Next, the underlying emotional intelligence components, along with the implementation of Gracie, the rapport-building system, are described. Following this is the experimental design and results. The dissertation concludes with a discussion of follow-on work.

Chapter 2

Related Work

This chapter reviews research on rapport-building agents, automated recognition of emotion, speech synthesis with emotion, and dialog systems that adapt to user emotional state.

2.1 Rapport and Nonverbal Behavior

A large amount of research has looked at building human-like relationships, such as rapport, in human-computer interactions. Reeves and Nass [28] have shown that people tend to interact with machines as if they were people. Knowing this, previous work has looked at how to achieve better interactions by focusing on nonverbal behaviors.

Many have suggested that in dyadic conversations the interlocutors influence each other, both in voice and in physical gesture [4]. Chartrand and Bargh [9] noted that timing is important: producing nonverbal behaviors such as backchanneling or interjecting at the correct time plays an important role in fluent speech. In particular, one theory that attempts to explain the rich interaction between humans independent of situations and context is Communication Accommodation Theory (CAT) [36]. CAT suggests that during interaction, people use verbal and nonverbal behaviors to establish social distance. The theory states that individuals "converge" their speaking styles in order to reduce social distance between interlocutors. People tend to converge by modifying nonverbal features such as speech rate, pauses, and utterance length, among others, during interaction. The theory suggests that convergence is used to gain approval or to engage in smoother interaction. There have been applications of this theory, for example, Lee et al. [22] have shown that during dyadic interaction these "influential" behaviors can be modeled and used to predict a user's state.

Others have built applications that provide nonverbal feedback to users by analyzing user state, captured through the speakers prosody. Previous work has shown that this can be done and that users like it for back-channel feedback [38] and appropriate prosodic variation [40] in acknowledgments. These demonstrations used hand-coded rules and worked in limited domains: smalltalk and trivial quiz dialogs respectively.

Embodied conversational agents (ECAs) are able to improve social relations with users by reacting with appropriate nonverbal behaviors. Cassell *et al.'s* [8] Rea, a virtual real estate agent, produced inter-utterance feedback when silence was detected in speech. Gratch et al. [16] implemented a virtual agent that produces nonverbal gestures. They created the Rapport Agent, a virtual agent that provided users with nonverbal human-like feedback. This was done by reacting to gestures, head position, gaze, voice intensity, and voice range, as indicators of back-channel opportunities. For example, if a back-channel opportunity arises and the agent is gazing forward, then the agent may nod its head. This approach was successful; the virtual human's nonverbal feedback positively affected speaker fluency and engagement during interaction: users produced fewer disfluencies and spoke longer with the responsive agent. One topic needing further study is integration with spoken content, which is essential in spoken dialog systems.

2.2 Emotion Recognition and Emotion Synthesis

Emotion research has had much attention in the past. This section reviews the two main approaches for describing emotions. It then reviews previous research in emotion recognition, including the emotions detected and the accuracies of the methods. Finally, it discusses research in emotion synthesis, along with methods and accuracies.

There are two main methods for describing emotion in voice. The first is a discrete approach. This describes emotions with English words, such as Eckman's six basic emotions: anger, disgust, fear, joy, sadness, surprise [12]. Another approach for describing emotions is using dimensions [30], typically two or three. These dimensions can be seen as components, each with its own value (usually a number from -100 to 100), that compose an overall emotional state. Three commonly used dimensions are valence (positive/negative), activation (active/passive), and power (dominant/submissive).

In the field of emotion recognition, most work has addressed the problem of recognizing discrete emotions. The best results have been realized with machine learning using mainly prosodic features (pitch, energy, speaking rate) and spectral features such as MFCCs [35]. At the end of the last decade, Petrushin et al. [25] achieved accuracies as high as 70% for five classes of emotion using a corpus of acted emotions. However, Batliner et al. [3] showed later that corpora containing acted emotions are different from spontaneous interactions. In particular, during spontaneous speech, humans tend to show more subtle emotions. Forbes-Riley and Litman's work [13] used the Adaboost algorithm (trained with a non-acted corpus) in a speech-only tutoring system to detect emotion. Their results were partly successful as they were able to detect three classes of emotions (positive, negative, and neutral) with up to 84% accuracy. Most recently, D'Mello et al. [11] attempted to detect a richer set of emotions including boredom, confusion, flow, frustration, surprise, and delight from conversation features (temporal information, response information, answer quality, tutor directness, tutor feedback). Detection of surprise and delight was not successful, but in general the system could detect the emotions with 54% accuracy.

More recent approaches use the dimensional approach to describe emotions. Many times emotion annotators find it difficult to label discrete emotions from voice; this may be due to the subtlety of emotions sometimes present in voice. With the dimensional approach, numerical values are usually assigned [22]. In Lee et al.'s work [22] utterances are labeled with two dimensions (activation and valence). Lee was able to recognize activation with 63% accuracy and valence with 65% accuracy even though the data had noise.

Regarding emotional speech synthesis, since Cahn's Affect Editor[7] it has been known that it is possible to generate human-detectable emotion with computers. The Affect Editor was capable of producing English sentences with different emotional colorings. Users were able to identify 44% of sentences with surprised emotional coloring (44% accuracy) and sadness with 91% accuracy. In the German language, as cited in [31], [5] was able to produce joy (81% accuracy) and fear (52% accuracy). Iida *et al.* [20] was able to use unit selection to produce anger, joy and sadness with good results (50-80% accuracies) in Japanese. These approaches generate discrete emotions with no variance in intensity. More recently, MaryTTS with the Emospeak interface [34], is capable of synthesizing emotional speech based on the three dimensional approach (see Figure 2.1). The emotion synthesizer was evaluated on activation and valence. Performance was measured by asking users to identify how well the "tone of voice" fit the emotion expressed in the words of different sentences. Correlations between the previously labeled emotions and the rated emotions from the study showed that activation and valence were perceivable [33].

2.3 Emotion-Adaptive Dialog Systems

Despite these findings, in spoken dialog systems, practical applications that use emotion to influence interaction are limited. One area of research where detection of uncertainty has proven to be useful is in the tutoring domain [14, 26]. Forbes-Riley et al.'s ITSPOKE tutoring system is capable of adapting responses to users based not only on the correctness of the answer, but also on how certain the user sounds. Users overall preferred the responses shown by the adaptive system and also had better learning efficiency (measured as learning gain divided by time on task and total student turns) [14].

Other domains require detection and production of different emotions. Klein et al. [21] showed that users are more willing to interact with a frustrating system longer if it shows signs of active listening, empathy and sympathy. Predinger et al. [27] developed a system that aimed at helping people prepare for interviews by adjusting to user emotional state, taken from physiological features. By providing empathic feedback, for example, when the user started feeling frustrated, the system was shown to reduce the user's stress. Similarly, Burkhardt et al. [6] extended a customer service-like dialog system to show empathy to

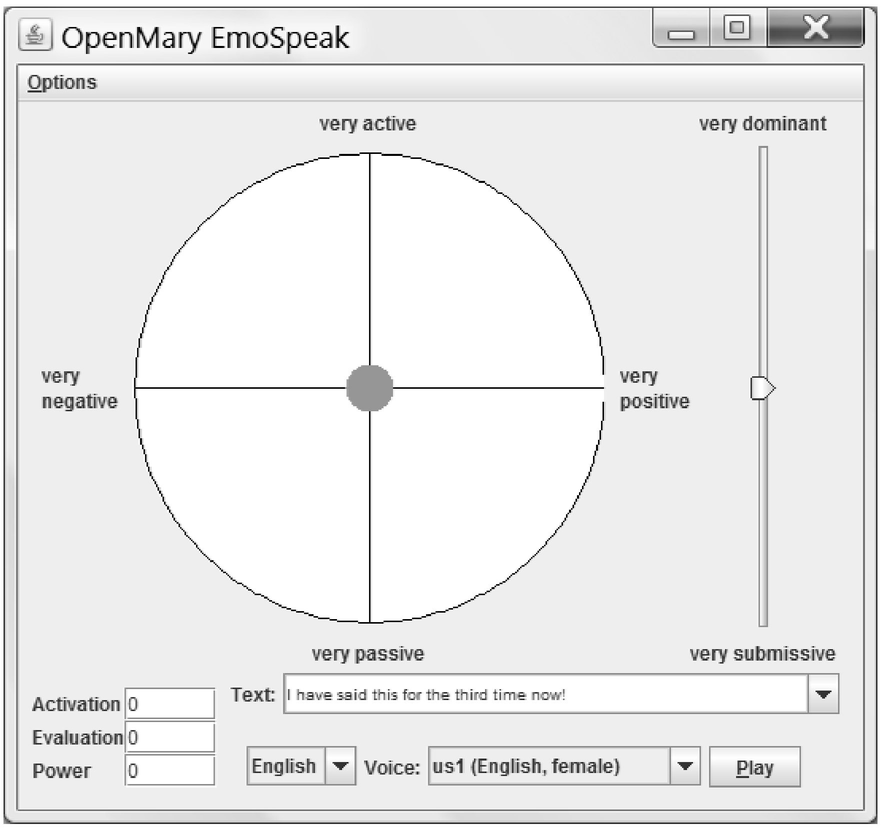

Figure 2.1: A snapshot of Emospeak, an interface for specifying emotional coloring for the MaryTTS speech synthesizer.

frustrated users. Depending on the number of speech recognition errors, the system would apologize and forward users to a human operator.

In the domain of persuasion, Portia [23] is a system that provides two types of argumentation content to users, depending on their responses. Rational arguments are aimed more at conflicts in reasoning (for example, if you are overweight, you are at a greater health risk, and you want to lower health risks) whereas emotional arguments are aimed at emotional factors (for example, if you eat too much, you may gain too much weight and you may lose your figure).

2.4 Summary

Regarding research in rapport building technologies, the CAT theory says that people use nonverbal feedback to establish social distance during conversation. In order to gain rapport, people would most likely want to decrease social distance in order to achieve the connectedness and smoothness in conversation that is seen in human social interaction. Research in human-computer interaction has pursued these nonverbal behaviors through appropriate back-channeling, head nods, and gaze techniques, and in very limited ways, dialog systems have begun to pay attention to user emotional state, which can be detected through some of these nonverbal behaviors in voice.

Emotion research has mainly focused on detection of discrete emotions, e.g., anger, disgust, fear, joy, sadness, surprise. These emotions have been difficult to detect, except when using a small set of emotions. This is probably because humans typically show more subtle emotions in most real human-human interactions [3]. The dimensional approach seems to be more accurate and more relevant, especially in spontaneous, non-acted, dialog. No systems that actually use dimensional emotions to interact with users have yet been built.

Once emotion is detected, the question that arises is how to react. The use of emotions for guiding interaction in dialog systems is limited. Some have used emotion to determine what content to provide to users, but none have attempted to emotionally color utterances at run-time to gain rapport with users.

In this dissertation I bring findings from previous work together and build a spoken dialog system that recognizes user emotional speech, using the dimensional method, and I show that by responding with appropriate emotional coloring, the dialog system is better at gaining rapport with users.

Chapter 3

The Role of Rapport in Persuasion

My interest in emotion and rapport grew as I attempted to create a better persuasive system. In this chapter, my study of persuasive dialog, which eventually led to the rapport building system, is described. First a corpus was collected and two prototype systems were built. Although the prototypes provided informative content for users, the interactions were unlike those seen in the corpus, and in particular emotional coloring. This chapter describes the corpus and then the two prototype systems. Some of the second prototype's design decisions were carried over to the final system.

3.1 Persuasiveness in Conversations about Graduate School

The starting point for this work was some preliminary data collection and analysis done before I took over the project. A persuasive dialog corpus was collected by Anais Rivera, a member of our lab. The corpus was not collected for the study of rapport or emotion. Rather it was collected with a vague intention of analyzing turn-taking cues, such as back-channels, in the domain of persuasion.

The corpus was collected by recording ten undergraduate students, who were enrolled in the introductory Computer Science course, speaking to the department's graduate coordinator.

The graduate coordinator's job functions included talking to undergraduates and helping to grow the graduate programs. This staff member was unusually personable and pleasant to talk to; thus we (among others) considered her an exemplar for effective dialog behaviors.

The ten students who spoke with the coordinator were compensated with credit for one of the assignments in their Introduction to Computer Science class. The students had little knowledge of the nature or value of graduate school and of the application process. The conversations lasted 9-20 minutes. For this research, 6 of the ten (2–7) dialogs were analyzed. Figure 3.1 shows a sample of an interaction.

Preliminary analysis of the corpus revealed that the dialogs were only mildly persuasive. There were no attempts to get the students to perform any immediate actions, and the dialogs were mostly about providing factual information, although of course the graduate school option was presented in a generally positive way. During the interactions with the students, it was clear that the coordinator had a set of possible topics she would cover during the interaction with the students. Examples of topics include financial issues, Graduate Record Exam information, and grade requirements. The coordinator's overall dialog strategy was to present topics based on some of the student's profile information like major, grade point average, and degree status (freshman, sophomore, junior, senior).

In an attempt to automate some of the persuasive behaviors seen in the corpus, a member of our lab, Rafael Escalante-Ruiz, created a baseline for a persuasive system. This baseline included a simple interface that would allow users to complete an electronic form that asked information that students in the corpus often revealed. This included interests, status (freshman, sophomore, etc.), GPA and concerns (financial, etc.). After the form was submitted, the data was passed into a program that concatenated text snippets to produce a letter customized for the user.

Figure 3.1 is an example letter. According to Rafael, users found the content informative. Some users wrote that the letter was more credible and required less effort than scanning information from a Website. However, it seemed there was a lack of persuasive impact. Part of the reason could be due to the textual nature of the content; the students reading it could just skim the material, disregarding some of the persuasive wording. In

Table 3.1: Excerpt from the persuasive dialog corpus.

Line	**Transcription**
GC0	How are you doing in your classes?
S1	*Good.*
GC1	Good, good.
S2	*Um, yeah good, good enough.*
GC2	Did they tell you why you're coming by to see me today?
S3	*No. Not really*
	...(conversation continues) ...
GC3	Have you ever thought about going to graduate school?
S4	*Yes.*
GC4	Yeah?
S5	*I have.*
GC5	Is this something that you probably want to do?
S6	*I would like to, yeah.*
GC6	You've got a long ways to go and a lot of time to think about it, cause this is your what? Second semester?
S7	*Yeah.*
GC7	Do you have any questions for me about it?
S8	*Well, I know that you have, don't you have to have a certain GPA to get into graduate school?*
GC8	Mhm. The GPA requirement is a 3.0.

comparison to the corpus dialogs, which had a coordinator speaking to students, the letter lacked interaction. Also, in the corpus, sometimes the coordinator would say information redundantly (probably to make the content more memorable); in the textual version, doing this would make the letter longer and maybe less appealing.

3.2 VoiceXML Prototype

Using what was learned from the letter generator baseline, a spoken dialog system was built in order to implement the type of interactions between the coordinator and the students seen in the corpus.

I built a VoiceXML prototype in collaboration with another graduate student, Jason Zheng. The system was made to behave like the coordinator in two critical respects: by delivering spoken content and being selective about the information that was provided.

3.2.1 Defining a Persona

Before building the dialog system, some basic design decisions were made in consideration of speech recognition limitations and natural language understanding limitations. Spoken dialog systems technology today is not able to handle some of the interactions in the corpus. For example, some of the human coordinator's responses were based on students' open ended answers. The coordinator could ask, *"Tell me why you are interested in graduate school,"* and if the student responded, *"Actually, I'm not, I don't really know much about it,"* the coordinator would go on to say, *"Really, well, I'm glad you came because now I can tell you more."* Limitations in speech recognition would make it difficult to handle this behavior; since the best results for speech recognition are when recognition is limited to a small set of words or word phrases. As the set of words increases, the accuracy of the recognizer decreases.

Another consideration is that when users interact with spoken dialog systems, even if the designer did not intend for it, users assign a personality to the system [10].

THE UNIVERSITY OF TEXAS AT EL PASO

Department of
Computer Science

Dear Natalia:

Thank you for using our Graduate School Advisor service. This letter summarizes some things to think about, based on the information you provided.

When you apply to graduate school, you will have to submit three things: the results of your GRE test, a statement of purpose and your grades. I will briefly explain each requirement in the following paragraphs.

The GRE is a structured examination designed to mainly test your verbal and mathematical skills. Since you already took the SAT, you have the experience of taking a structured test: you can expect the same mechanics in the GRE. One difference that you will find is that the GRE is computer-based, and it may require you to adapt to the computer interface. To ensure that you score as high as you can, it is essential to prepare for the GRE. Buy one of the review books and develop a systematic plan that will enable you to brush up on your skills in vocabulary, reading comprehension, analogies, algebra, and geometry. From the GPA that you provided, I can see that you are a good student, and you should not have trouble getting a good score in the GRE.

You are unique and thus have different goals than other students. Your grades and your GRE scores do not reflect these goals and your motivation for attending graduate school; this is why the admissions committee asks for a written Statement of Purpose as part of your application. The Statement of Purpose is an essay written by you by which you tell the committee about your life objectives and what you expect to gain from a graduate degree. This allows the committee to have a better grasp of you as a person, instead of forming a mental picture of you from numbers alone.

One fact that you might not be aware of is that the graduate committee favors students with GPAs above 3.0. Your major grades appear to be good, but the overall grades that you reported are below this favored point average; however, you are still a freshman and have several semesters ahead of you in which you can improve your GPA. The committee will place more weight on the later semesters when they evaluate you, so, if you decide you want to attend graduate school, don't get discouraged! This is an opportunity for change and the right time to begin to earn excellent grades.

One of the advantages of being a graduate student, and specially a PhD student, is that there are a lot of different ways to pay for your education without taking money out of your pocket. There are scholarships, fellowships, assistantships and on-campus jobs, where people prefer to hire a graduate student because the assumption is that the student is more mature and has a greater body of knowledge to draw from. If you apply for the PhD program and we accept you, we almost guarantee that we will find money to pay you. Actually in our department there is not a single PhD student who is not getting funding in some way or another. As you can see, the outlook for graduate school looks a lot different from paying for your undergraduate degree.

If you want to discuss Graduate School personally, I would be delighted to talk to you.

Sincerely,

Rafael Escalante

Rafael Escalante
UTEP Graduate School Student

El Paso, Texas
79968-0518
(915) 747-5480

Figure 3.1: An automatically generated customized letter, built using content from the persuasive dialog corpus.

To reduce the impact of these technical limitations and to influence users' perception of the dialog system's personality (to make it like the human coordinator's, and yet appropriate for what can actually be implemented), a persona was created. When defining a persona it is important to describe some background information. For this system, the persona is described in Figure 3.2.

> Sophia is 38 years old and has been an employee at the University for ten years. She grew up in El Paso, Texas and is familiar with the variety of cultures in the area. Being an extrovert, Sophia is outspoken and always very enthusiastic when it comes to higher education. Sophia enjoys working with students because she was once a student herself. She received her PhD degree at the University and has since been involved in many activities to promote higher education. Sophia is determined to encourage students to attend graduate school just as she once was. During her undergraduate studies, she was determined to graduate and start working. Sophia never thought that she would ever go back to school.
>
> Things changed during Sophia's senior year in her degree when one of her teachers told her about the great possibilities possible with a Masters or PhD degree. Sophia eventually finished her undergraduate degree and started working. However, when working she found that her career goals had not been met. She gained an interest in research and eventually finished her PhD degree. She is now working as a researcher in the department of Computer Science and loves her job. She has decided that one of her main purposes in life will be to share information about the opportunities of graduate school with students.

Figure 3.2: Persona background information

In summary, the words that the persona uses are meant to exhibit

- someone who likes to talk to people
- a person who understands student academic situations
- someone who will not be evaluating or judging towards students
- a trustworthy person
- a friendly person
- a helpful person
- an approachable person regardless of a student's background
- an expert about the graduate program, the application process, and requirements
- an expert about funding opportunities

The persona was used to guide us during the process of selecting content to include and also for choosing substitution phrases when the speech synthesizer did not correctly pronounce the original words from the corpus.

3.2.2 Adaptive Content

The prototype was built to see what aspects of persuasion would be missing in a system built using current commercial technology. The prototype attempts to motivate users of the system to attend graduate school by adapting the content given to users based on their degree status and interests, as the coordinator did.

To develop an effective plan of advice, the system first asks the student about their degree status. Based on this, the system selects content accordingly. As the dialog progresses, the student is asked about interests in other topics in order to continue the planning. This is done in a system-directed way. In particular, the majority of questions are menus and yes/no questions. Occasionally, open ended questions are asked, but the student's answers are ignored in order to avoid speech recognition errors. Barge-in is disabled.

Table 3.2 shows an interaction with a junior. Table 3.3 shows an interaction with a senior. Annotations indicate how some topics are enabled and disabled depending on the students' answers. The choice of content differs in the two dialogs. One difference is that with the senior, the system immediately gives more details about the graduate school experience (i.e. research), but with the junior, is less direct.

One important note is that the information presented by the prototype system is without any semantic or pragmatic prosodic variation, there are only small adjustments of prosody to suit the words and syntax present, done automatically by the speech synthesizer in order to make the synthesized utterance sound more natural: for example, the first word of an utterance is produced relatively slowly, and, if the words form a question, the speech will have a slightly positive pitch slope at the end of the utterance.

3.2.3 Spoken Dialog System Architecture

Most spoken dialog systems are built with a similar architecture (see Figure 3.3). When the user speaks, the sound is captured and passed to the speech recognizer. The speech recognizer passes word hypotheses to the dialog manager. The dialog manager decides, based on the user's words, what information should be given next. The speech synthesizer then speaks to the user.

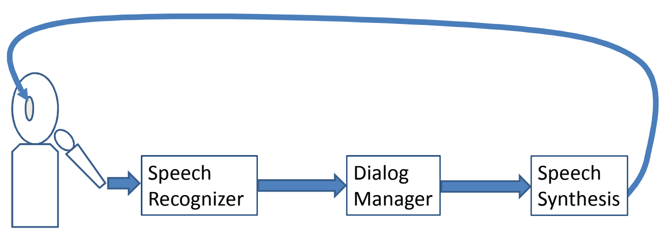

Figure 3.3: Traditional spoken dialog system architecture.

Of the three components seen in Figure 3.3, the prototype system uses the speech recognizer and speech synthesizer provided by Nuance Cafe [1]. This commercial VoiceXML

Table 3.2: Sample interaction with between the VoiceXML prototype and a student with a junior classification

Line	Transcription
SYS1	Hi I'm Sophia. Are you interested in graduate school?
USR2	*Yes.* [‘enables’ graduate school requirements]
SYS3	Are you a freshman, sophomore, junior or senior?
USR4	*Junior.* [‘enables’ advice for juniors]
SYS5	System: What's your name?
USR6	*Jaime.* [answer is ignored, static acknowledgment is given]
SYS7	Nice name. Is your major computer science?
USR8	*Yes.* [‘enables’ computer science class information]
SYS9	Are you finding your computer science classes to be hard?
USR10	*No.* [‘disables’ advice for improving grades and for getting help]
SYS11	OK.
SYS12	A rising junior. Well, that's good. So in maybe a couple years you would have probably come by to see me anyway because that's when people really start thinking about “oh my gosh i'm getting ready to graduate what am i gonna do now?”
SYS13	There are just a few requirements for the graduate school, would you like to hear about them?
USR14	*Yes.* [‘enables’ graduate school requirement information]
SYS15	I think our University's graduate school asks that you have at least a 3.0 in order to get into the graduate school.
SYS16	Would you like me to tell you some tips on how to write a great statement of purpose?
USR17	*No.* [‘disables’ statement of purpose information]
...(conversation continues) ...	
SYS18	Thanks for calling, bye.

Table 3.3: Sample interaction with between the VoiceXML prototype and a student with a senior classification

Line	Transcription
...(starts the same as the previous example through SYS11, then continues)...	
SYS1	There are many choices after you graduate with your bachelor's degree. Would you like to hear about a few of them?
USR2	*Yes.* ['enables' career choice information]
SYS3	One thing that makes graduate school different is that there's a huge research component. If you're getting your master's you take classes just like you would as an undergraduate. But at the end of it you write a thesis, and, in order to write that thesis you've got to pick a question you've got to go out and do research on how to answer that question.
SYS4	After you graduate you can either work right away, or you can continue into graduate school. One of the most common reasons for continuing to graduate school is that students graduated and they couldn't find a job. hopefully that's not gonna happen in your case. If teaching is your long term goal, well, you can teach at a community college with your master's. And if you want to teach at a university, then you need your PhD.
SYS5	I'm always available, I can help you a lot in your future. Would you like to hear about scholarships or fellowships?
USR6	*Yes.* ['enables' scholarship and fellowship information]
SYS7	I have some scholarships that are accepting applications right now. For example, ... (conversation continues) ...
SYS8	Well let me tell you this, I want you to keep in touch. I'm here all the time. Thanks for calling, bye.

hosting network allows developers to build voice applications by uploading VoiceXML code to a voice server. The code is executed when a user calls a particular telephone number. This service is free for non-commercial use.

3.3 Dialog Manager

Since VoiceXML lacks the ability to create a dialog manager able to model the strategic behavior seen in the corpus (real time enabling and disabling of topics and digressing to topics based on user interest), the third component, the dialog manager, was made from scratch. The dialog manager consists of three components. First, to generate VoiceXML files for the voice server in real time, the Generator is needed. Second, the Topic Database containing all of the dialog content is needed. Third, to keep track of what topics have been used and which topics should be used in the future, the Maintainer takes the words the user speaks and updates the database.

The Generator is a PHP program that takes as input three text segments: a prompt, a VoiceXML grammar (which specifies the possible words and word sequences that can be recognized by the speech recognizer), and an acknowledgment (something that is produced as a quick response to the user before the next topic is said). The Generator then generates an VoiceXML file. This VoiceXML file is then rendered by the voice server, using speech synthesis to say the prompt and acknowledgment. Speech recognition is used to interpret the user word response.

The Maintainer is a PHP program that takes as input a user response (e.g., yes/no) and updates the database by enabling and disabling topics. The Maintainer also queries the database for enabled topics and puts them into a local stack. The local stack contains topics that will be used the future. The Maintainer also passes the next topic (the one at the top of the stack) to the Generator so that it can create a VoiceXML file.

The following examples show how information from the Topic Database, is tranformed into dialog. This occurs when the Generator takes this information, generates a VoiceXML file, and then passes it to the voice server. The Maintainer processes the user response and updates the Topic Database accordingly.

- **Example 1**

  Given the turn specification:

  > **TopicContent:** So, what's your name?
  >
  > **Grammar Type:** *Record*
  >
  > **Acknowledgement:** Nice name.
  >
  > **Maintainer Action:** (none)

  This is expanded by the Generator to produce VoiceXML, which then allows exchanges like:

  > System: So, what's your name?
  >
  > User: Jaime
  >
  > System: Nice name.

- **Example 2**

  Given the turn specification:

  > **Topic Content:** Is your major computer science?
  >
  > **Grammar Type:** *Boolean*
  >
  > **Acknowledgment:** (none)
  >
  > **Maintainer Action:**
  >
  > if (response == yes)
  >
  > ‘enable’ Computer Science info, ‘disable’ non-CS info, etc...
  >
  > if (response == no)
  >
  > ‘enable’ non-CS info, ‘disable’ Computer Science info

  This is expanded by the Generator to produce VoiceXML, which then allows exchanges like:

System: Is your major Computer Science?

User: yes

System: How are you liking your Computer Science classes?

- **Example 3**

  Given the turn specification:

  Topic Content: Are you a freshman, sophomore, junior or senior?

  Grammar Type: *Choice*

  Acknowledgement: (none)

  Maintainer Action:

  if (response == freshman)

  'enable' advice for grades, 'disable' current scholarship info

  if (response == sophomore)

  'enable' advice for grades, 'enable' research groups info

  if (response == junior)

  'enable' advice for jobs, 'enable' Masters degree info

  if (response == senior)

  'enable' PhD and Masters info, 'enable' current scholarship info, 'disable' advice for grades

  This is expanded by the Generator to produce VoiceXML, which then allows exchanges like:

  System: Are you a freshman, sophomore, junior or senior?

  User: sophomore

  System: I think our University's graduate school asks that you have at least a 3.0 in order to get into the graduate school.

The third component of the dialog manager is the topic database. The database contains the list of topics that can be spoken by the spoken dialog system (see Appendix A). It contains rules for choosing future topics (for example, if the user stated an interest in teaching, then, in addition to teaching information, PhD degree information will also be given in the future). Lastly, it keeps track of which topics have already been used in the current dialog. The following are the fields in the Topic Database's "dialogs" table.

1. TopicContent - the prompt that is associated with the topic (e.g. "Hello, I'm Sophia").
2. Grammar - the set of possible words that can be recognized from the user's speech. There are three types of supported grammars. The *record* type is meant for open ended answers. The user response is not processed, but when the end of utterance is detected, the system responds with an acknowledgement. The *boolean* grammar tells the system that the user will answer either yes or no. Finally, the *choice* grammar allows for menus.
3. Acknowledgment - An immediate response that the system speaks after receiving user input data.
4. Enabled - a boolean field that tells whether this topic should be used in the future.
5. Completed - a boolean field that tells whether the topic has already been used. If set to true, the topic will not be subsequently re-enabled.
6. TopicID - a numeric unique identifier for each topic.
7. TopicName - the name of the topic (e.g. systemGiveGreeting).
8. Section - Each topic belongs to a section. Related topics are in the same section. This number contains the section that the topic is associated with (e.g. the topics *systemGiveAdviceForFuture* and *systemGiveAdviceForFutureGrades* are in the section of *Encouragement*).
9. Internal order - within a section, the internal order number for each topic, specifying the order that the system cover the topics.

3.3.1 Prototype Control Flow

The following describes the flow of control between the components of the prototype system. See Figure 3.4 for a pictorial representation.

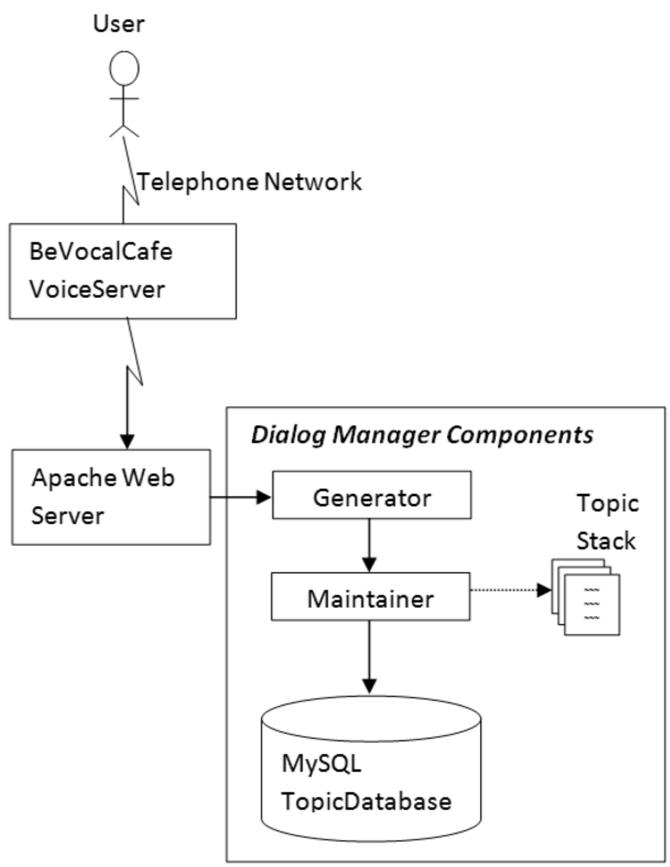

Figure 3.4: Prototype system controlflow diagram.

When a user calls the Nuance BeVocalCafe voice server using a telephone, the following occurs:

1. The Generator calls the Maintainer and alerts it that a new session has begun
2. The Maintainer then initializes the database enabled/completed fields. At this point only topics for greeting and acquiring user classification are enabled.

3. Based on the user's answers, the Maintainer updates the database by enabling and disabling certain topics (see Appendix A for the specifics).

4. Next, the Maintainer queries the database for enabled topics and stores the results into a local stack. The local stack is a cache for topics that will be used; this reduces the number of queries to the database. If any of the disabled topics are in the local stack, they are removed.

5. In the Maintainer, an entry from the local stack is popped and passed to the Generator.

6. The Generator creates a VoiceXML file from the entry.

7. The generated VoiceXML file is then rendered by the voice server and the user's speech is recognized.

8. The recognized speech is then passed to the Maintainer.

9. This process repeats until the local stack is empty and there are no more enabled entries in the database.

3.3.2 Prompt Tuning

After the system was implemented, there were some issues concerning the speech synthesizer, such as prompts being too long and some words being difficult to understand by human listeners. This was fixed by introducing pauses and using different words. Especially in long sentences, the text-to-speech system seemed to do a poor job of pausing and pronouncing words. In addition, in the corpus, the graduate coordinator used words like gonna, wanna, hafta, and also disfluencies. These often sounded unnatural, so these were replaced with traditional English words (going to, want to, ...).

3.4 Lessons Learned from the Prototype

Building the VoiceXML prototype suggested that using spoken dialog systems for persuasion may be a good decision. They seemed to be closer to human-human persuasion than the simple letters described at the beginning of this chapter. Unlike the letter version, in the dialog system, the information was provided in smaller chunks, making it easier to process. Also, the information was spoken; users could not skim. Since the information was given using speech (which is slower than reading text), the dwell time for each topic was longer; this probably made the information more memorable. Also, users were able to partially choose the topics covered during the conversation by showing interest or disinterest during the dialog. The VoiceXML prototype was not the final solution; there were still many differences between the human coordinator in the corpus and the system, especially regarding the prosody in the interaction. To learn more, several users were asked for their opinions.

Four laboratory members interacted with the system. They were asked to jot down their impressions of shortcomings during their interaction with the prototype (Appendix C contains the questionnaires). Most of the subjects stated problems widely known to afflict spoken dialog systems: factors related to speech recognition, end of utterance detection delays, and lack of mixed-initiative interaction. More interestingly, subjects stated that the coordinator sounded bored and uninterested. In contrast, when listening to the corpus, it was clear that the coordinator did not sound bored at all. The coordinator seemed to vary her prosody including pitch, timing, and volume. This did not appear to be accidental or random, rather it seemed to be the primary way that the coordinator showed attention, involvement, and empathy. The dialogs with the VoiceXML systems lacked this, and the overall impression was very different; it was as if the users were browsing a collection of audio clips, rather than having a real interaction.

Based on this, the direction of the research was confirmed: to investigate how humans produce emotions in voice by varying their prosody during social interaction, and especially in a persuasive setting where the development of rapport is essential. The next chapter describes this work.

Chapter 4

Aspects of a Model of Human Emotional Interplay

The aim of this work is to build a system capable of gaining rapport with users. In order to accomplish this, a corpus containing spontaneous persuasive dialog was analyzed for content and structure, as described in the previous chapter.

This chapter describes the annotation of the corpus with emotions as well as the development of an Emotion Recognizer. Next, the responsive strategies found in the corpus are described along with the learning of a set of rules for predicting what emotional stance a persuasive system should take in response to the user's inferred emotional state.

4.1 Identifying Emotions in the Corpus

Although the collection of the corpus was intented to analyze turn-taking behavior in persuasive dialog, I found, listening to the corpus, there were other noticeable nonverbal signals present between the interlocutors. This was heard in prosodic variations in voice. It seemed the coordinator was emotionally coloring her responses based on the student's emotional state in the immediately preceeding utterance. I set out to analyze this. With the creation of a spoken dialog system that could respond to a user's utterance in mind, and considering the need for fairly conservative turn-taking, the first step in the analysis was to segment the corpus into a set of utterances. For this purpose, I defined an utterance as a segment of speech that starts when a speaker begins a turn and ends when the other speaker either interjects or begins a turn. Speech in times of overlap is also treated as separate utterances.

Initially, selected utterances in the corpus were annotated using categorical emotions (e.g., sad, happy, angry, disgusted, scared, and surprise). There was little agreement between the judges annotating the corpus. A possible reason for this was that the emotions displayed in the corpus were subtle. This is to be expected; previous work has shown that in spontaneous speech, versus acted speech, humans do not exhibit extreme emotions [3]. Therefore, a different approach was taken for describing the emotions in the corpus, namely, the dimensional approach [32].

Emotion labels were assigned to each utterance. Two judges were asked to label utterances on the three dimensions. The judges worked independently. To avoid distraction when switching between two speakers, each judge labeled emotions in the dialogs one side at a time. First they labeled all of the coordinator's utterances, then they labeled all of the student's utterances. Before the judges annotated each dialog in the corpus, they were asked to listen to a random list of utterances of the speaker, so that they could become accustomed to the speaker's normal speaking styles.

The judges were asked to listen to each utterance three times. The first time they assigned a value (–100 to +100) for activation, the second time for valence, and the third time for power. The following definitions were given to the judges:

- Activation (+100=Extremely Active,–100=Extremely Passive) If a speaker is active, it sounds like he/she is engaged and shows interest in his or her voice. A passive voice would sound like a lack of engagement or interest.

- Valence (+100=Extremely Positive,–100=Extremely Negative) This dimension represents the sound of pleasure in the voice. Positive may be shown by sounding upbeat or pleasant, whereas negative may sound down or unpleased.

- Power (+100=Extremely Dominant,–100=Extremely Submissive) A dominant sounding voice can sound like the speaker is taking control or is very sure of what he/she is saying. A submissive voice sometimes sounds like there is uncertainty or like he/she is trying to not show too much power in voice.

Table 4.1: Correlation coefficients between the two judges' annotations of emotions in the persuasive dialog corpus

Emotional Dimension	**Inter-judge Correlation**
Activation	0.58
Valence	0.42
Power	0.62

The annotations for each dimension were correlated between the two judges. Table 4.1 shows the correlation coefficients. Spearman correlations were used due to the non-normal distribution of the annotations. The possible values for correlation coefficients range from [0,1]. General meanings for correlation coefficients are the following: 0-0.3 are low, 0.3-0.5 are medium, and 0.5+ are high. One reason may be due to problems with some of the utterances in the corpus.

Specifically, agreement was lower for utterances that are very short (such as backchannels or grunts), or that have too much noise (from the speakers moving their microphones). Also, sometimes within an utterance, the speaker's prosody changed drastically; this means that the speaker's may have had emotional shifts during an utterance. The judges were asked to assign one emotion labeling for the entire utterance.

Table 4.2 shows an excerpt of the corpus with emotion labels, noting some salient acoustic properties.

Table 4.2: Annotated excerpt from the persuasive dialog corpus

Line	**Transcription**	**Emotion (Act., Val., Pow.)**	**Notable Acoustics**
GC0	So you're in the 1401 class?	(35, 10, 35)	normal speed, articulating beginnings of words
S1	*Yeah.*	*(10, 5, –5)*	*higher pitch*
GC1	Yeah? How are you liking it so far?	(40, 10, 35)	medium speed, articulating beginnings of words
S2	*Um, it's alright, it's just the labs are kind of difficult sometimes, they can, they give like long stuff.*	*(5, -10, -15)*	*slower speed, falling pitch*
GC2	Mm. Are the TAs helping you?	(20, –10, 10)	lower pitch, slower speed
S3	*Yeah.*	*(5, 5, –15)*	*rising pitch*
GC3	Yeah.	(20, 5, –15)	rising pitch
S4	*They're doing a good job.*	*(10, 0, 5)*	*normal speed, normal pitch*
GC4	Good, that's good, that's good.	(35, 10, 40)	normal pitch, normal speed

4.2 Building an Emotion Recognizer

In order to test whether a spoken dialog system with emotional intelligence is capable of building rapport, several components were needed (see Figure 4.1). Some components were available as open source, namely the Speech Recognizer and the Emotional Speech Synthesis. The dialog manager was described in the previous chapter. The other components had to be built based on analysis of the interactions in the corpus. This section describes the steps taken to build the initial version of the Emotion Recognizer component.

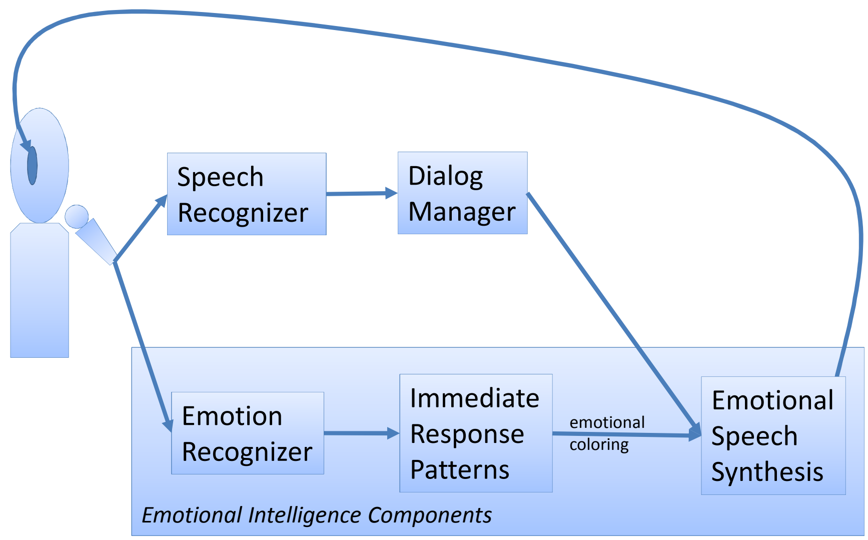

Figure 4.1: A high level view of the architecture of a spoken dialog system with emotional intelligence. The main system components and their connections are shown.

The Emotion Recognizer was built based on one of the judge's annotations of the utterances in the corpus. To build the Emotion Recognizer, features from the utterances (pitch, energy, and timing) were extracted and then a recognizer was trained.

Three software tools were used to extract features from the recorded utterances. A prosody analysis tool in our lab, Dede, was used to extract mainly pitch and energy features.

Many of the pitch and energy features used were selected because they are important indicators of emotion in voice as seen in previous research [22]. There were also features that were used because they have proven to be useful in determining turn-taking and backchannel cues [37, 29, 39, 38]. Although there was no good reason to think these would be useful for inferring any particular emotion, I left them in and, in fact some of these features seemed to help slightly in the recognition of emotion. MRate [24] estimated the speaking rate of the speaker from an utterance. The Hidden Markov Toolkit (HTK) [42] was used to extract mel-frequency cepstral coefficients. For each tool, although the features were calculated every 10ms, only the averages over the length of the utterance were used for training the recognizer. Also, all features were computed for each utterance in isolation, without normalization relative to each speaker's averages. Tables 4.3 - 4.5 show the features used.

Table 4.3: Acoustic features used for training the Emotion Recognizer (basic features)

Feature	Description
zerox	number of zero crossings in the current 10ms frame; if low, indicates a vowel or voiced consonant.
xcvs	indicator based on zero crossings (0 for consonant, 1 for vowel, 2 for silence)
period energy	energy over the 10ms frame
log energy	log of the energy over the 10ms frame
energy slope	slope over the previous 40ms, useful for detecting back-channel cues in Arabic
delta	absolute values of the difference between current pitch point and previous pitch point (10ms earlier) divided by the current pitch
vowels seen	the number of vowels in the utterance so far

Table 4.4: Acoustic features used for training the Emotion Recognizer (energy related features).

Feature	Description
max log energy	maximum log energy
energy average	average energy
average vowel vol	average energy inside the vowels
ecvs	indicator based on energy (0 for consonant, 1 for vowel, 2 for silence)

An emotion recognizer was trained using WEKA [41]. Specifically, the M5P [15] algorithm was trained with the acoustic features taken from the feature file written by Dede, HTK, and MRate. These acoustic features were used to predict the value on each emotion dimension as labeled by the judge. M5P was used for two reasons. First the trained model was simple (a set of linear equations). This would make integration with a dialog system easier. Also, M5P performed best among the various linear regression models.

For training, ten-fold cross validation was used. This means that the file was split into ten pieces. Nine of the ten pieces were used for training while the remaining piece was used to evaluate the performance of the trained model. All of the speaker data was mixed, (i.e. the test data was not always an unseen speaker). Figures 4.2 - 4.4 show the model to calculate the value for the activation dimension. Figure 4.5 shows the model to calculate the value for the valence dimension. Figures 4.6 - 4.8 show the model to calculate the value for the power dimension.

Table 4.7 contains the correlations between the predicated values (from the trained model) and the actual values (the judge's original label) for each dimension of emotion. Correlation coefficients for activation and power are high, but moderate for valence. Future work could consider adding other acoustic features or using a different machine learning

Table 4.5: Acoustic features used for training the Emotion Recognizer (pitch related features).

Feature	Description
avg pitch	average pitch
median pitch	a fixed percentile of the pitch, in this case 28%
decile1 pitch	10th percentile of the pitch
decile9 pitch	90th percentile of the pitch
local med pitch	a local variable representing an estimate of the median pitch
local median pitch	an array storing values for the above pitch references
pitch status	1 if the pitch is valid (typically a vowel or voiced consonant); 0 if the frame is unvoiced (silence or a voiceless consonant)

Table 4.6: Acoustic features used for training the Emotion Recognizer (composite features traditionally used for detecting back-channel opportunities)

Feature	Description
stridency	a measure of how much the pitch is going harshly up and down, at one point thought to be useful for Arabic
flat	boolean, true if the current frame is part of a region of flat pitch, useful for English and Japanese
downslope len	if there is an ongoing monotonic drop in pitch, how long it has lasted, useful for Arabic
highness len	if the pitch has been relatively high for a while, how long it has been, useful for Spanish
lowness len	if the pitch has been relatively low for a while, how long it has been, useful for Spanish, English, and Japanese
avg vowel len	the length of the average vowel seen so far, at one time thought to be useful for Spanish
prevous utterance len	duration of the previous utterance, useful for English, Spanish, Arabic, and Japanese
edginess, fast edginess	measures of how much the signal currently seems to be a decisive drop in energy marking a turn end
edge d, fast edge d	booleans indicating whether the current frame is part of an edgy region

Table 4.7: Correlation coefficients (used to measure quality) between the output of the Emotion Recognizer and the labeled emotional values.

Emotional Dimension	**Recognizer Quality (Correlation Coefficient)**
Activation	0.73
Valence	0.44
Power	0.79

algorithm in order to increase the performance of the Emotion Recognizer.

Some further analysis showed that using portions of the utterance (for example, the last 500ms) were good indicators of emotion. However, the results were not as good as when using the entire utterance. In addition, when leaving out HTK and Mrate parameters, the correlation coefficients decreased only slightly. For this reason, for the final system only Dede features were used. With the availablility of the newly developed openSMILE package [2] developed by TUM as part of the SEMAINE project, future work can take advantage of this 300+ feature extraction system.

4.3 Immediate Response Patterns

The next step was to determine how to model the way that the coordinator emotionally colors her responses. This section describes the empirical analysis used to do this, and ultimately to build the Immediate Response Patterns component.

4.3.1 Adjacency Pairs

To determine whether the coordinator was reacting to students' emotional states, each student utterance was grouped with one coordinator utterance and the two were considered

```
local_median_pitch <= 145.21 :
|   lowness_len <= 0.145 : ACT_RULE1
|   lowness_len >  0.145 : ACT_RULE2
local_median_pitch >  145.21 :
|   xcvs <= 1.855 :
|   |   prev_utterance_len <= 745.785 : ACT_RULE3
|   |   prev_utterance_len >  745.785 :
|   |   |   avg_vowel_vol <= 6.48 : ACT_RULE4
|   |   |   avg_vowel_vol >  6.48 : ACT_RULE5
|   xcvs >  1.855 : ACT_RULE6
```

ACT_RULE1:

```
activation =
      + 64.0653 * edge_d
      + 28.3514 * pitch_status
      + 31.4256 * xcvs
      - 9.8611  * log_energy
      + 7.7723  * avg_vowel_vol
      + 7.2245  * max_log_energy
      + 1.1057  * fast_edge_d
      + 0.5216  * ecvs
      - 0.5101  * vowels_seen
      + 0.4298  * downslope_len
      + 0.3055  * lowness_len
      + 0.1302  * edginess
      + 0.1168  * decile9_pitch
      - 0.0921  * fast_edginess
      - 0.078   * highness_len
      - 0.0554  * local_median_pitch
      + 0.0356  * median_pitch
      - 0.0203  * period_energy
      + 0.0167  * prev_utterance_len
      + 0.0116  * avg_pitch
      - 0.0022  * decile1_pitch
      - 0.0008  * avg_vowel_len
      - 0.0005  * highest
      - 170.9463
```

ACT_RULE2:

```
activation =
      - 107.3701 * edge_d
      - 61.9757  * vv_ratio
      + 50.2428  * fast_edge_d
      + 49.1969  * xcvs
      + 18.3251  * avg_vowel_vol
      - 12.1029  * max_log_energy
      - 3.3652   * highness_len
      + 2.3725   * pitch_status
      - 1.6104   * local_median_pitch
      + 1.0746   * median_pitch
      + 0.6901   * ecvs
      + 0.4298   * downslope_len
      + 0.368    * lowness_len
      + 0.3249   * avg_pitch
      + 0.23     * log_energy
      + 0.195    * vowels_seen
      + 0.0346   * edginess
      - 0.0272   * avg_vowel_len
      - 0.0203   * period_energy
      - 0.0054   * fast_edginess
      + 0.0037   * decile9_pitch
      - 0.0022   * decile1_pitch
      + 0.0007   * prev_utterance_len
      - 0.0005   * highest
      + 44.1125
```

Figure 4.2: Calculation of activation from acoustic feature values.

ACT_RULE3:

```
activation =
      - 83.5352 * fast_edge_d
      + 76.0695 * pitch_status
      + 26.3199 * log_energy
      - 23.8536 * zerox
      - 15.1935 * max_log_energy
      - 14.6531 * energy_average
      + 11.0641 * downslope_len
      - 10.1024 * ecvs
      + 7.2856  * avg_vowel_vol
      + 6.7361  * xcvs
      + 6.0947  * edge_d
      + 0.164   * lowness_len
      + 0.0925  * median_pitch
      - 0.0876  * avg_pitch
      - 0.0772  * avg_vowel_len
      - 0.0264  * highness_len
      + 0.0235  * vowels_seen
      + 0.0228  * decile9_pitch
      - 0.0104  * fast_edginess
      - 0.0057  * period_energy
      - 0.0033  * decile1_pitch
      + 0.0032  * edginess
      - 0.0037  * highest
      + 0.0031  * local_median_pitch
      + 0.0002  * prev_utterance_len
      + 91.97
```

ACT_RULE4:

```
activation =
      - 22.2155 * fast_edge_d
      + 11.4628 * downslope_len
      - 10.4391 * ecvs
      + 6.7361  * xcvs
      - 6.3873  * log_energy
      + 6.0947  * edge_d
      + 5.3437  * avg_vowel_vol
      + 3.395   * pitch_status
      - 1.5927  * max_log_energy
      - 1.2343  * zerox
      + 1.1882  * energy_average
      - 0.8263  * highness_len
      + 0.164   * lowness_len
      + 0.0965  * median_pitch
      - 0.0917  * avg_pitch
      + 0.089   * period_energy
      + 0.0421  * vowels_seen
      + 0.0237  * decile9_pitch
      - 0.0106  * fast_edginess
      - 0.0037  * highest
      - 0.0033  * decile1_pitch
      + 0.0032  * edginess
      + 0.0031  * local_median_pitch
      - 0.0004  * avg_vowel_len
      + 0.0002  * prev_utterance_len
      + 52.8915
```

Figure 4.3: Calculation of activation from acoustic feature values (continued).

ACT_RULE5:

```
activation =
      - 41.8957 * pitch_status
      - 22.2155 * fast_edge_d
      - 10.4391 * ecvs
      + 11.4628 * downslope_len
      + 6.7361  * xcvs
      - 6.3873  * log_energy
      + 6.0947  * edge_d
      + 5.2837  * energy_average
      + 3.634   * avg_vowel_vol
      - 1.5927  * max_log_energy
      - 1.4532  * highness_len
      - 1.2343  * zerox
      + 0.164   * lowness_len
      + 0.0965  * median_pitch
      - 0.0917  * avg_pitch
      + 0.0561  * vowels_seen
      + 0.0237  * decile9_pitch
      - 0.0106  * fast_edginess
      - 0.0057  * period_energy
      - 0.0037  * highest
      - 0.0033  * decile1_pitch
      + 0.0032  * edginess
      + 0.0031  * local_median_pitch
      - 0.0004  * avg_vowel_len
      + 0.0002  * prev_utterance_len
      + 67.2351
```

ACT_RULE6:

```
activation =
      + 41.362  * pitch_status
      - 23.9436 * zerox
      + 15.3558 * log_energy
      - 14.4838 * ecvs
      + 14.0822 * max_log_energy
      - 5.2628  * fast_edge_d
      + 4.9326  * xcvs
      + 3.9748  * edge_d
      + 1.7136  * downslope_len
      + 1.4615  * avg_vowel_vol
      + 0.4716  * local_median_pitch
      - 0.4098  * median_pitch
      + 0.164   * lowness_len
      - 0.0488  * period_energy
      - 0.0419  * highest
      - 0.0264  * highness_len
      - 0.0222  * edginess
      + 0.0046  * vowels_seen
      - 0.0043  * fast_edginess
      + 0.0036  * avg_pitch
      - 0.0033  * decile1_pitch
      + 0.0018  * decile9_pitch
      + 0.0002  * prev_utterance_len
      - 0.0004  * avg_vowel_len
      - 50.137
```

Figure 4.4: Calculation of activation from acoustic feature values (continued).

```
decile9_pitch <= 182.5 : VAL_RULE1
decile9_pitch >  182.5 : VAL_RULE2
```

VAL_RULE1:

```
valence =
      + 11.2611 * fast_edge_d
      + 9.9792  * xcvs
      + 8.0299  * ecvs
      + 6.0248  * downslope_len
      + 4.0257  * delta
      + 2.2412  * lowness_len
      + 2.2316  * max_log_energy
      + 1.9676  * avg_vowel_vol
      + 1.0218  * energy_average
      - 0.9851  * highness_len
      + 0.8055  * edge_d
      + 0.7129  * vv_ratio
      + 0.3093  * log_energy
      - 0.1625  * local_median_pitch
      - 0.1127  * decile9_pitch
      + 0.0907  * decile1_pitch
      + 0.0889  * median_pitch
      + 0.017   * edginess
      - 0.0016  * avg_pitch
      - 0.0011  * fast_edginess
      - 0.0007  * period_energy
      + 0.0007  * highest
      - 62.0891
```

VAL_RULE2:

```
valence =
      + 42.964  * edge_d
      + 31.8211 * vv_ratio
      + 23.673  * log_energy
      + 25.2998 * xcvs
      - 20.7619 * fast_edge_d
      - 6.5871  * zerox
      + 4.2021  * avg_vowel_vol
      - 4.1228  * delta
      + 0.2458  * median_pitch
      + 0.1792  * downslope_len
      - 0.1755  * local_median_pitch
      + 0.1201  * ecvs
      - 0.0319  * period_energy
      + 0.0302  * highest
      - 0.0273  * fast_edginess
      + 0.0201  * edginess
      - 0.0016  * decile1_pitch
      - 0.0011  * avg_pitch
      + 0.0011  * decile9_pitch
      - 236.1008
```

Figure 4.5: Calculation of valence from acoustic feature values.

```
local_median_pitch <= 152.83 :
|   avg_pitch <= 93.045 : POW_RULE1
|   avg_pitch >  93.045 :
|   |   period_energy <= 90.73 : POW_RULE 2
|   |   period_energy >  90.73 : POW_RULE 3
local_median_pitch >  152.83 :
|   zerox <= 2.305 :
|   |   highness_len <= 2.615 : POW_RULE 4
|   |   highness_len >  2.615 : POW_RULE 5
|   zerox >  2.305 : POW_RULE 6
```

POW_RULE1:

```
power =
      - 136.188 * energy_slope
      + 52.079  * edge_d
      - 49.8944 * vv_ratio
      + 24.1502 * xcvs
      + 15.3161 * ecvs
      + 3.8455  * pitch_status
      + 0.8413  * avg_vowel_vol
      + 0.4632  * lowness_len
      + 0.337   * downslope_len
      - 0.3305  * fast_edge_d
      - 0.2427  * log_energy
      + 0.1389  * decile9_pitch
      + 0.1363  * edginess
      - 0.108   * fast_edginess
      - 0.0427  * energy_average
      - 0.0265  * decile1_pitch
      + 0.0241  * prev_utterance_len
      + 0.0224  * median_pitch
      + 0.0141  * vowels_seen
      + 0.0057  * local_median_pitch
      - 0.0022  * period_energy
      - 0.002   * avg_pitch
      - 63.1037
```

POW_RULE2:

```
power =
      + 76.0704 * edge_d
      - 13.9691 * vv_ratio
      + 7.366   * xcvs
      - 4.9391  * fast_edge_d
      + 4.8896  * avg_vowel_vol
      + 3.4901  * pitch_status
      - 3.2334  * ecvs
      - 2.1521  * log_energy
      + 1.5342  * lowness_len
      + 0.6221  * period_energy
      - 0.5195  * highness_len
      + 0.337   * downslope_len
      - 0.148   * decile1_pitch
      + 0.0615  * median_pitch
      + 0.0576  * vowels_seen
      - 0.0427  * energy_average
      + 0.0312  * avg_pitch
      + 0.0202  * edginess
      - 0.0119  * fast_edginess
      + 0.0059  * decile9_pitch
      + 0.0057  * local_median_pitch
      + 0.0005  * prev_utterance_len
      - 68.2882
```

Figure 4.6: Calculation of power from acoustic feature values.

POW_RULE3:

```
Power =
      + 38.0068 * xcvs
      - 22.3562 * pitch_status
      - 21.0031 * max_log_energy
      + 16.25   * avg_vowel_vol
      + 8.164   * lowness_len
      + 7.4433  * edge_d
      - 4.2582  * vv_ratio
      - 1.7354  * fast_edge_d
      - 0.8017  * log_energy
      + 0.2749  * avg_pitch
      - 0.4892  * decile1_pitch
      - 0.9856  * ecvs
      + 0.337   * downslope_len
      + 0.2596  * vowels_seen
      - 0.1584  * highness_len
      - 0.0527  * decile9_pitch
      - 0.0427  * energy_average
      + 0.0326  * median_pitch
      + 0.0109  * edginess
      - 0.0067  * fast_edginess
      + 0.0057  * local_median_pitch
      - 0.0042  * period_energy
      + 0.0005  * prev_utterance_len
      + 36.1051
```

POW_RULE4:

```
power =
      + 97.1935 * xcvs
      + 46.9128 * pitch_status
      - 43.5224 * zerox
      - 25.4642 * edge_d
      - 6.2755  * fast_edge_d
      + 4.9888  * log_energy
      + 3.4198  * downslope_len
      + 3.246   * vv_ratio
      + 1.1576  * delta
      - 0.594   * energy_average
      + 0.442   * avg_vowel_vol
      + 0.2516  * lowness_len
      - 0.1106  * median_pitch
      - 0.0326  * period_energy
      - 0.0247  * avg_pitch
      + 0.0197  * edginess
      + 0.0153  * fast_edginess
      + 0.0092  * decile9_pitch
      + 0.0092  * local_median_pitch
      + 0.0054  * vowels_seen
      - 0.0039  * highest
      + 0.0003  * prev_utterance_len
      - 83.9913
```

Figure 4.7: Calculation of power from acoustic feature values (continued).

POW_RULE5:

```
power =
      + 45.3883 * xcvs
      + 26.8531 * pitch_status
      - 25.4642 * edge_d
      + 18.3361 * delta
      + 16.4954 * max_log_energy
      - 6.2755  * fast_edge_d
      - 5.914   * zerox
      + 4.9888  * log_energy
      + 3.4198  * downslope_len
      + 3.246   * vv_ratio
      - 0.594   * energy_average
      + 0.442   * avg_vowel_vol
      - 0.3103  * median_pitch
      + 0.2516  * lowness_len
      - 0.0859  * period_energy
      - 0.0247  * avg_pitch
      + 0.0153  * fast_edginess
      + 0.0092  * decile9_pitch
      + 0.0092  * local_median_pitch
      - 0.0069  * edginess
      + 0.0054  * vowels_seen
      - 0.0039  * highest
      + 0.0003  * prev_utterance_len
      - 107.1458
```

POW_RULE6:

```
power =
      - 190.6763 * edge_d
      + 69.6079  * pitch_status
      + 21.1011  * log_energy
      - 13.4994  * zerox
      + 11.8023  * max_log_energy
      - 9.6809   * ecvs
      + 4.1284   * xcvs
      - 3.5306   * energy_average
      - 3.0882   * fast_edge_d
      + 1.8225   * downslope_len
      + 1.4453   * vv_ratio
      + 0.5154   * delta
      + 0.442    * avg_vowel_vol
      - 0.1291   * edginess
      + 0.2516   * lowness_len
      + 0.0819   * fast_edginess
      - 0.0278   * highest
      - 0.0189   * period_energy
      - 0.0127   * avg_pitch
      + 0.0092   * local_median_pitch
      + 0.0068   * median_pitch
      + 0.0059   * decile9_pitch
      + 0.0039   * vowels_seen
      + 0.0003   * prev_utterance_len
      - 59.3932
```

Figure 4.8: Calculation of power from acoustic feature values (continued).

an adjacency pair. In the normal case, an adjacency pair consisted of an utterance by the student and a subsequent response by the coordinator. As a special case, if both spoke simultaneously, the coordinator's utterance was treated as a response to the student's, as it seemed that her adaptation was fast enough for there to sometimes be a causal relation even in such cases. In total there were 962 adjacency pairs across the six dialogs.

4.3.2 Correlations between the Interlocutors' Emotions

I hypothesized that there were "immediate response patterns" determining how the coordinator chose an emotional coloring for her utterance. That is, in response the emotion expressed by the student in the immediately previous utterance. This hypothesis was based on Communication Accommodation Theory (as mentioned in chapter 2); a matching of the nonverbal features between two interlocutors to decrease social distance. In this case, it was expected that the student's emotion, as expressed through nonverbal prosodic variation, would be mirrored in the coordinator's response. Correlations across the adjacency pairs were measured; the results are given in Table 4.8.

Table 4.8: Correlation coefficients between coordinator emotion dimensions and subject emotion dimensions in adjacency pairs.

		Student		
		Activation	Valence	Power
Coordinator	Activation	**–0.14**	0.14	–0.24
"	Valence	0.04	**0.34**	-0.05
"	Power	–0.15	0.12	**–0.31**

In the valence dimension there was clear evidence for mirroring: the correlation coefficient was 0.34. This makes sense: if the student is positive about something the coordinator will tend to take that perspective, and similarly for negative feelings. For example, in Table

4.9 in adjacency pair S2-GC2, the subject speaks slower and with a falling pitch (which sounds negative) and the coordinator (GC2) mirrors his negative voice. Of course this pattern does not mean that the coordinator slavishly mimicked the student's attitudes, however it was common for her to at least acknowledge his feelings before going on. For example, in response to a student who expressed a negative attitude about the financial burdens of graduate school, she first acknowledged that money was a serious concern, in a sombre voice, but in subsequent utterances turned positive as she explained the opportunities for funding.

In the power dimension there was an inverse relationship, a –0.31 correlation: if the student sounded dominant, the coordinator generally became more submissive and vice versa. This was probably mostly a reflection of the natural give-and-take of a dialog: when one person is taking the floor, the other person is yielding it. For example, in adjacency pairs in Table 4.9 GC0-S1 and GC1-S2 the coordinator is clearly leading the conversation and the student following. This pattern also is not invariable; in S3-GC4 it appears that the student's *yeah* is submissive in the sense that he wants to say no more on this topic, but the coordinator thwarts him by also disclaiming any attempt to take the floor, forcing him to make a more explicit statement in S4.

In the activation dimension the picture is less clear; again there was a negative correlation, but a much weaker one. In fact, the coordinator's activation seems to relate more to the student's power: as the student sounds more dominant, the coordinator becomes more disengaged (–0.24 correlation).

4.3.3 Building a Predictive Model

These results indicate that the coordinator is executing some emotionally responsive strategies during these dialogs. While it is interesting to examine such strategies, as done above, ultimately the aim was to build a system to determine appropriate responses, and for this machine learning of appropriate rules holds more promise than a labor-intensive study of specific strategies.

Table 4.9: Annotated excerpt from the persuasive dialog corpus.

Line	**Transcription**	**Emotion (Act., Val., Pow.)**	**Notable Acoustics**
GC0	So you're in the 1401 class?	(35, 10, 35)	normal speed, articulating beginnings of words
S1	*Yeah.*	*(10, 5, –5)*	*higher pitch*
GC1	Yeah? How are you liking it so far?	(40, 10, 35)	medium speed, articulating beginnings of words
S2	*Um, it's alright, it's just the labs are kind of difficult sometimes, they can, they give like long stuff.*	*(5, -10, -15)*	*slower speed, falling pitch*
GC2	Mm. Are the TAs helping you?	(20, –10, 10)	lower pitch, slower speed
S3	*Yeah.*	*(5, 5, –15)*	*rising pitch*
GC3	Yeah.	(20, 5, –15)	rising pitch
S4	*They're doing a good job.*	*(10, 0, 5)*	*normal speed, normal pitch*
GC4	Good, that's good, that's good.	(35, 10, 40)	normal pitch, normal speed

Thus machine learning methods were applied in attempts to build a predictor for the coordinator's emotional responses observed in the adjacency pairs. The students three emotion dimensions were taken as attributes and were used to predict the coordinators emotional coloring, again using only the annotations by the first judge. Machine learning algorithms from WEKA were used, with ten-fold cross validation, and I measured the correlations between the predictions of the model and the actual values in the corpus. Among the algorithms tested were the MultilayerPerceptron, SVM, Linear Regression, and Tree-based models like M5PTree and REPTree. The best performing algorithms were REPTree and Bagging with REPTree [41]. This may be due to the fact that the judge's emotion labels were usually on values divisible by 5. Table 4.3.3 shows the results and Figures 4.9 - 4.11 show the models for calculating values for activation, valence, and power respectively.

Table 4.10: Correlation coefficients between actual dimension value and predicted dimension value using student dimension levels as attributes, with the highest correlations in bold

Student Dimensions	**Predicted Coordinator Dimensions**	**Prediction Correlations**	
		REPTree	**Bagging**
Act, Val, Dom	Act	**0.24**	0.19
Val, Dom	Val	0.28	**0.35**
Act, Val	Dom	**0.34**	0.30

Overall the results show that it is possible to predict, to some extent, the emotional coloring to use based only on the emotions expressed in the previous utterance.

Wondering why the results were not higher, the absolute errors for each dialog were averaged. The first dialog (which was collected before the others) had the highest average absolute errors in all dimensions. The student in this first dialog seemed to have a distinct speaking style (West Coast accent and persistently creaky voice); another likely factor was

```
activation response value=
      SubjAct < 7.5
      |   SubjVal < -30 : 36.33
      |   SubjVal >= -30
      |   |   SubjPow< -2.5
      |   |   |   SubjAct < -7.5 : 62.6
      |   |   |   SubjAct >= -7.5 : 71.21
      |   |   SubjPow>= -2.5
      |   |   |   SubjAct < -7.5 : -10
      |   |   |   SubjAct >= -7.5 : 60.49
      SubjAct >= 7.5
      |   SubjVal < 32.5 : 45.86
      |   SubjVal >= 32.5
      |   |   SubjAct < 45
      |   |   |   SubjPow< 22.5
      |   |   |   |   SubjAct < 15
      |   |   |   |   |   SubjPow< 2.5 : 30
      |   |   |   |   |   SubjPow>= 2.5 : 71.67
```

Figure 4.9: Calculation of systems activation level based on the human subject's emotion in the immediately preceding utterance.

```
valence response value =
      SubjVal < 55
      |   SubjVal < -30 : -11.93
      |   SubjVal >= -30
      |   |   SubjPow < -12.5 : 9.12
      |   |   SubjPow >= -12.5
      |   |   |   SubjPow < 22.5 : 19.21
      |   |   |   SubjPow >= 22.5
      |   |   |   |   SubjPow < 77.5
      |   |   |   |   |   SubjAct < 55 : 10.28
      |   |   |   |   |   SubjAct >= 55 : -2.23
      |   |   |   |   SubjPow >= 77.5 : 31.88
```

Figure 4.10: Calculation of systems valence level based on the human subject's emotion in the immediately preceding utterance.

```
power response value =
      SubjAct < 7.5
      |   SubjVal < -7.5 : 27.47
      |   SubjVal >= -7.5 : 63.93
      SubjAct >= 7.5
      |   SubjAct < 77.5
      |   |   SubjAct < 45 : 45.77
      |   |   SubjAct >= 45
      |   |   |   SubjVal < 27.5 : 17.84
      |   |   |   SubjVal >= 27.5 : 64.02
      |   SubjAct >= 77.5 : -5.2
```

Figure 4.11: Calculation of systems power level based on the human subject's emotion in the immediately preceding utterance.

that the persuader was probably still devising her strategies during this first meeting.

Other reasons may be due to problems with the recorded audio. Across all speakers, one characteristic of the worst predicted coordinator responses was poor recording quality, when one of the interlocutors was fidgeting, or was too far from the microphone or otherwise sounded muffled. In addition, overlapping utterances and short utterances (less than one second) were common in the poorly predicted cases. On the other hand, listening to the best predicted pairs, the utterances were generally longer, clearer and not overlapping.

4.4 Summary

The results presented in this chapter provide a foundation for the system described in the next chapter. There are two key findings. First, when people speak, they color their utterances with emotions that are subtle, but still perceivable by the listener. Second, the emotional coloring in the utterances is influenced by the listener's perceived emotional state.

Chapter 5

Implementing Gracie, the Rapport System

This chapter describes the implementation of Gracie (GRAduate Coordinator with Immediate-response Emotions). To recap, the main question of this dissertation is, can rapport (which may be essential for persuasion) be gained with a spoken dialog system that produces emotionally appropriate responses? This chapter is ordered as follows. First the technologies used for the implementation are described. Next, the components that make up Gracie are presented. Lastly, the system dataflow is explained.

5.1 Choice of Implementation Platform

To facilitate the development of Gracie, several software tools were used. These included an operating system where the tools could be run and an integrated development environment (IDE) for compiling the different components into a single project. In addition, I needed a programming language that supports publicly available third party libraries for sound recording/playing, speech recognition, and others.

Gracie runs on the CentOS distribution of the Linux operating system. This is because several open source packages for recording and playing sound, speech recognition, speech synthesis and others are available for this platform. In addition, since it is open source, it can be installed onto any computer, without the need to purchase a license. Also, some software that is used in Gracie (previously built in our lab) is also written for Linux. This avoided having to port these tools to another operating system.

Eclipse was chosen as the IDE for Gracie mainly because it is available for free. In addition, it is continuously maintained and well documented. Eclipse allows users to import projects (in Gracie there were several projects) and with little effort, compile them into a single executable file. Lastly, Eclipse was chosen because of my familiarity and experience with it.

To implement the novel components of the system, C++ was used. This is because all of the other components were written in either C or C++ (the only exception is the speech synthesizer, but this was not a problem as discussed in section 5.2.3). Also, C++ is supported by the Linux operating system as well as by Eclipse.

5.2 System Components

Gracie was built by integrating several open source components and some newly built components. Figure 5.1 shows the overall system architecture. The Emotion Recognizer and Immediate Response Patterns are based on findings in chapter 4. The Dialog Manager is based on the VoiceXML prototype version. The Speech Recognition and Emotional Speech Synthesis components are open source and publicly available. This section will describe these components in more detail.

5.2.1 Immediate Response Patterns

To give Gracie the ability to adapt emotional speech to the user's emotional state, the trained machine learning algorithms described in chapter 4 are implemented as components in the system. The trained models are produced in WEKA and implemented in Gracie.

The Emotion Recognizer was a set of three models (one model for each emotion dimension) trained using an M5P linear regression tree. Each model takes as input a set of extracted features from a sound signal. These features are extracted using Dede. The combined output from the three models is the emotional state of the user using the three-dimensional representation.

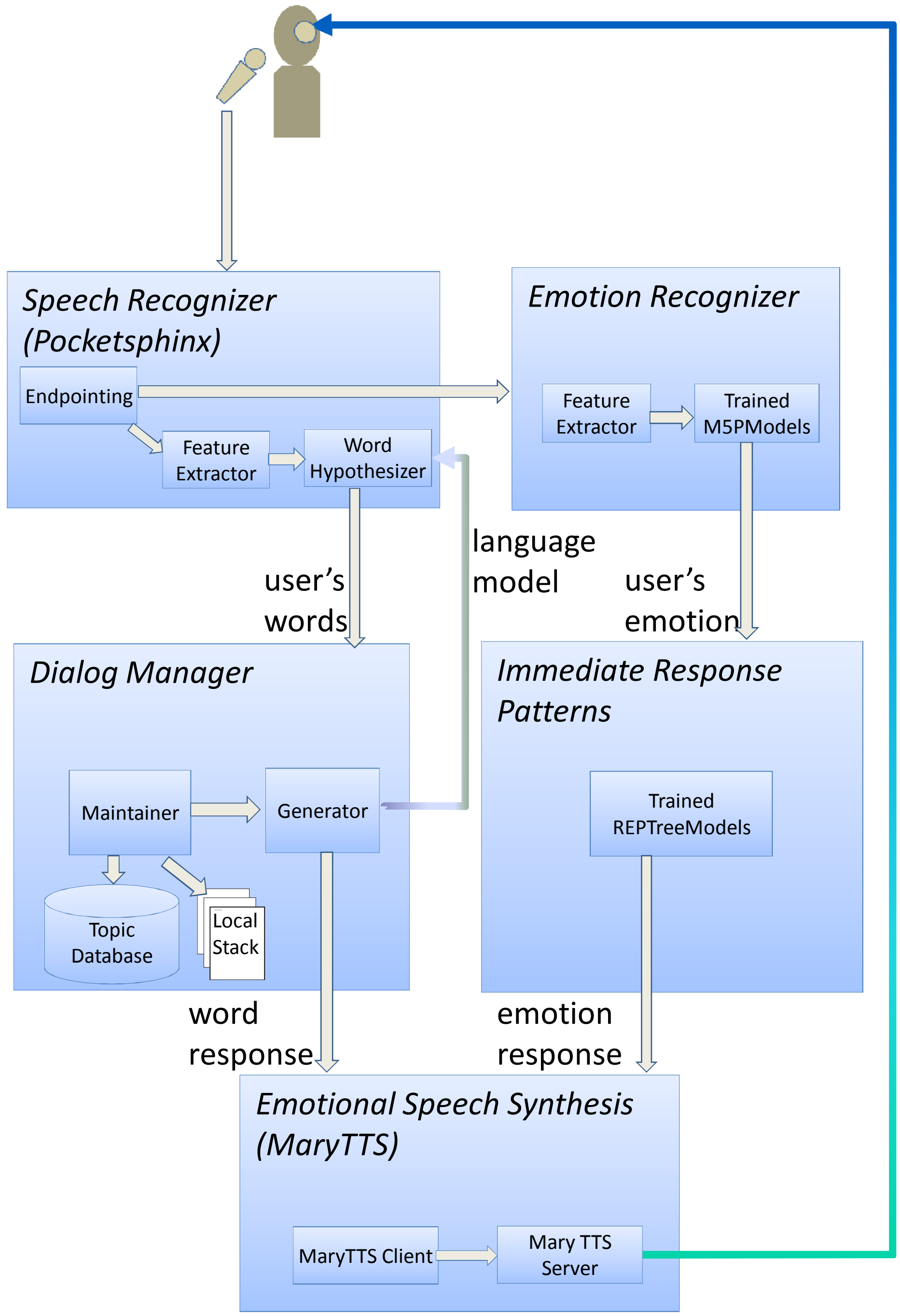

Figure 5.1: Architecture of Gracie

In order to implement the Emotion Recognizer in Gracie, the user's voice must be processed in real-time (during or immediately after the user speaks, instead of processing off-line). This is to avoid latency in responses to users. Aizula, a sister program of Dede, is capable of doing this. Aizula extracts features from the sound signal. This provides the input to the Emotion Recognizer's trained models. Since the models are linear regression trees, they are a collection of "if" statements and simple equations. This made coding the model in C++ trivial. The models for calculating emotion are described in section 4.3.3.

The output from the Emotion Recognizer, the perceived three-dimensional emotion, is then passed to the Immediate Response Patterns component.

The Immediate Response Patterns component takes this three-dimensional representation of emotion and provides an appropriate three-dimensional emotional response. Similar to the Emotion Recognizer, the Immediate Response Patterns component is a set of three models (one model for each emotion dimension). The linear regression algorithm used is REPTree. The model is a set of "if" statements. The resulting emotion triple is used to emotionally color the voice of the dialog system, which is done using the Emotional Speech Synthesizer, MaryTTS as described below. The formulas were described in section 4.3.3.

5.2.2 Dialog Manager

The dialog manager for Gracie behaves the same as in the VoiceXML prototype: it presents topics using the strategies in Appendix A. This means that the Topic Database remains unchanged; both systems use the same data. The implementation of the other two components in the VoiceXML dialog manager (see Figure 3.4) needed slight modification.

In the VoiceXML prototype, the Generator formatted data (into VoiceXML) for use by the speech synthesizer and speech recognizer. Both the speech synthesizer and speech recognizer were part of the BeVocal voice server. The voice server only took VoiceXML as input. In Gracie, the Generator has the same task, but is more simple. Now it provides the words that are spoken by the speech synthesizer and the language model for the speech recognizer without the VoiceXML tags.

The Maintainer's task in Gracie is the same as in the VoiceXML prototype. It queries the database for the next topics; store these in a local stack; and also updates the Topic Database depending on the words that the user speaks. The only difference is that instead of being implemented using PHP, it is implemented in C++ for easier integration with the other components in Gracie.

5.2.3 Emotional Speech Synthesizer

The interaction in spoken dialog systems consists of listening to the user, and also responding with words. The word responses are produced by a speech synthesizer. Traditionally, speech synthesizers only slightly tailor the intonation of the spoken content to suit the words and syntax present. In the case of Gracie, in order to produce emotions through voice, a synthesizer that allows modifying the prosody of its output is needed. MaryTTS [34] is such an application.

In addition to its prosodic flexibility, MaryTTS has an extension named Emospeak. With Emospeak, a three-dimensional emotion triple is entered along with some words as input; the output is an emotionally colored utterance. MaryTTS and Emospeak are implemented in Java with a client-server architecture. The MaryTTS server has a socket interface (which is language independent) that allows clients to connect using a TCP/IP interface. The client sends the words that will be spoken, as well as the values for prosodic variations. After processing, the server returns an audio file consisting of the speech. The client then renders the audio file. For Gracie, a client was built in C++.

5.2.4 Speech Recognizer

In spoken dialog systems, the dialog progresses based on the users word responses. In order to determine the words that are spoken by the user in response to questions from the dialog system, a speech recognizer is needed.

For Gracie, the speech recognizer is not as important as in most dialog systems. The possible words that are recognized by Gracie are few, and the main concern is to respond to emotions from voice, not the words spoken. The main factors in choosing a Speech Recognizer are recognition speed; accuracy comes second. Also, an open source solution is important because of the need for close integration with Aizula (again for speed reasons).

PocketSphinx [19] is used in Gracie for several reasons. First, it is open source. Second, since PocketSphinx is meant to operate in embedded systems, it does not require a top-of-the-line processor. Third, although it may not have accuracies as high as more comprehensive recognizers, its recognition speed, which is within two seconds on a Pentium M 1.3Ghz processor, is adequate for Gracie. This includes a one second pause necessary for end-of-utterance detection. The one second pause was chosen after testing Gracie with users.

5.3 Dataflow

The following example shows how the components of Gracie exchange data.

1. After Gracie greets the user, the user speaks his or her name

2. The user's response is passed to the Endpointing component, which determines the beginning and end of the user's utterance and trims the response. The trimmed response is then passed to the Speech Recognizer's Feature Extraction component and to Emotion Recognizer's Feature Extraction component.

   As in the prototype,

   (a) The Speech Recognizer extracts acoustic features from the utterance and recognizes a word or word sequence.

   (b) The words are then sent to the Maintainer, which enables and disables certain topics in the database (based on the user's words).

(c) If the local stack is not empty, then the Maintainer will pop off the next topic and pass it to the Generator.

(d) Otherwise, the database is queried for enabled topics; these are placed on the local stack. The Maintainer will then pop off the next topic and pass it to the Generator.

(e) The Generator will then pass the topicContent (the words that will be spoken) to the MaryTTS client.

Unlike the prototype,

(a) In parallel, the Aizula component extracts features from the user's utterance and passes them to the Trained M5P models.

(b) The M5P models convert the acoustic features into values for emotion dimensions.

(c) The values for the emotion dimensions are then passed into the REPTree models and a different set of emotion dimensions are sent to the MaryTTS client.

3. MaryTTS takes both the words and emotion values and generates an audio file. The audio file is sent to the MaryTTS client and played back to the user.

4. The system keeps the same emotional coloring for its utterances until the Immediate Response Patterns generate a new emotion, which happens after the next user input.

After the implementation of the system, the next step of the research was to determine its value. Human users were asked to evaluate Gracie in terms of its ability to gain rapport as described in chapter 6.

Chapter 6

Experimental Design

In order to determine the ability of Gracie to gain rapport, a user study was conducted that involved users interacting with three versions of the system. This chapter describes the experimental configurations of Gracie, the experimental procedure, and the subjects involved in the study.

6.1 Conditions

The research question of this dissertation is about the value of emotionally intelligent responses in voice. To answer the research question, whether a spoken dialog system with emotional intelligence if better at gaining rapport than a spoken dialog system without emotional intelligence, some changes were made to Gracie.

Gracie was configured to three versions. To determine whether Gracie was better at gaining rapport with users, two controls were used: a Neutral and a Non-Contingent version. The Neutral version would allow comparison between a system with appropriate emotional responses and a system with no emotional responses. The Non-Contingent version was included to determine the difference between a system with appropriate emotional responses and emotional responses not necessarily based on the user's state. A similar method was used by Gratch et al. in their studies on rapport [18].

The following are the three versions of Gracie used for the experiment.

- Neutral - The users emotional state is ignored. The utterances are spoken without emotional coloring. This version interacts with users in the same way as current dialog systems.

- Rule-based - The system recognizes the user's emotional state from voice, calculates an appropriate emotional response (based on the findings in chapter 4), and speaks back to the user with appropriate emotional coloring.

- Non-Contingent - Based on Gratch et al.'s method [18]. In this case, the users emotional states are ignored. Utterances are colored, but the colorings are read from a file, not calculated at runtime. The emotions in the file are the sequence of triples calculated by the rule-based system during the interaction with the previous user.

Each version of Gracie may be used with one of three different content sequences. The first is advice about grades (contentA), the second is about the statement of purpose (contentB), and the third is about the Graduate Record Examination (contentC). In total there are 9 possible configurations of Gracie.

6.2 Simplifications for Robustness and Usability

To avoid errors due to speech recognition, which may affect user's ratings for the systems, the Topic Database grammars were changed. Before, three types of grammar types were allowed (*boolean*, *choice*, and *record*). Both *boolean* and *choice* have the dialog manager produce content depending on the words spoken by the user. For the experiment, only the acoustic features in the voice are needed to determine how to color a response based the user's emotional state. For this reason, Gracie only takes free response answers, and thus ignores the user's words. The Topic Database was accordingly changed to only contain *Record* grammars.

Table 6.1: Fixed dialog used for the evaluation of Gracie (ContentA).

Topic Content	Acknowledgment
Hi, I'm Gracie, tell me your name.	Nice to meet you. I'm really glad you called. Now I can tell you about graduate school. There are a couple of things before you get into graduate school. It is a good idea to keep your grades up.
Graduate school is a great choice. Getting an advanced degree will help you get interesting jobs. It helps avoid boring work. Tell me what you think is a boring job.	Interesting answer.
We have a lot of our graduate students get great jobs when they finish. They do research at great companies. Microsoft and Google are examples. These companies look for people with higher degrees. Tell me what you know about graduate school.	Yeah, graduate school opens many doors.
Tell me why you chose computer science.	Oh I see. That is very interesting. I changed my major three times.
Our University is great. We have good teachers. We also have a lot of research groups. Tell me what you know about research groups.	Alright.
We have research groups in several topics. For example, the Interactive Systems Group looks at how humans interact with computers.	Thanks for listening.

Table 6.2: Fixed dialog used for the evaluation of Gracie (ContentB).

Topic Content	Acknowledgment
Hi, I'm Gracie, tell me your name.	Nice to meet you. I'm really glad you called. Now I can tell you about graduate school. There are a couple of things before you get into graduate school. You have to write a statement of purpose.
The statement of purpose is a short essay. It helps you express your reason for wanting a graduate degree. Tell me about your writing skills.	OK. I always ask others to proofread my work.
There are a lot of differences in graduate school. One is funding. You can do research and get paid. This way you also have job experience when you graduate. Tell me about your work experience.	It's good to have experience before looking for a job.
Tell me what you think of the University.	The school is expanding very quickly. We have a lot of funding too.
A lot of people are choosing graduate school. Some people do it because they can't find a job. A good reason to go on is if you have a career goal. Tell me your career goals.	Alright. That's good.
When you graduate you get paid more with a masters degree. You get about ninety thousand as a Doctor.	Thanks for listening.

Table 6.3: Fixed dialog used for the evaluation of Gracie (ContentC).

Topic Content	Acknowledgment
Hi, I'm Gracie, tell me your name.	Nice to meet you. I'm really glad you called. Now I can tell you about graduate school. There are a couple of things before you get into graduate school. You have to take an exam.
The exam is not hard. It is about the same as a high school exam. There is a reading, math, and writing part. It is worth getting into graduate school. Tell me what you think of exams.	Interesting answer.
Each teacher does research. You can ask each one about their work. Graduate school is about expanding the field of Computer Science. Tell me about your favorite computer science class.	Good.
Tell me how your teachers and graders help you do better in your classes.	Oh I see. That is very interesting.
If you want to teach you will need a graduate degree. Usually a University requires a Doctoral degree to teach. Tell me your plans for your future.	Alright, that's good.
It takes longer to get a graduate degree, but you will enjoy it. You meet great people and do fun research.	Thanks for listening.

The Topic Database was also changed to reduce the length of the interactions. In pilot studies the average interaction with Gracie was roughly 8 minutes. Since users in the study interact with three versions of Gracie, this would make for long interactions, where users may lose focus and interest. In addition, users would listen to the same (or very similar) speech during each interaction. To improve this, a different content sequence is used for each interaction (Tables 6.1 - 6.3). This is to keep users interested and focused throughout the three interactions. Each set of content makes for roughly 5 minute interactions with users, depending on how much the user talks.

During pilot studies, some subjects noted that the speech in Gracie was difficult to understand. This was especially true when the utterances were colored with high values of activation, which are spoken faster. For this reason, the words that Gracie spoke were simultaneously presented textually on the terminal for users to read if they chose.

Lastly, sometimes during interactions, subjects would begin talking before the system would start recording. This caused problems for the emotion recognizer because the feature extraction component would have a partial utterance to process. This problem was fixed by displaying the words "Please Speak" when the system was ready to record.

6.3 Improvements based on User Comments

After making these changes, and running several people in the lab, the experiment was started. After completing 36 subjects, the results were analyzed, and, contrary to expectation, the results showed that there were no significant preferences for any version of Gracie. Specifically, the majority of users preferred the neutral version. Fortunately, the cause was obvious after reading the subjects' comments. The majority commented that the rule-based and non-contingent versions spoke either too fast or too slow. The speaking rate was improved by retraining the Immediate Response Patterns component with normalized emotion labels. The normalization technique used was z-normalization. After normalization, the values ranged from -3 to +3. They were then multiplied by 10 to lie within -30 to +30. The trained model, therefore, produced values in the same range.

Another problem was that the rules for determining Gracie's emotion responses had large discontinuities. A slight change in the user's emotion would cause Gracie to respond with a drastic change. One user mentioned that Gracie seemed "bi-polar" during the interaction. The issue with discontinuities was resolved by retraining the Immediate Response Patterns component. Instead of using REPTree, Linear Regression from WEKA was used. The new rules are in Figure 6.1. The new system architecture is in Figure 6.2. With the new model, Gracie's responses seemed more natural to people in the lab talking to it. Thus, the comments provided by users suggested that implementing small improvements would improve the ability of Gracie to gain rapport.

activation response value = 0.20SubjVal - 0.16SubjPow + 6.03

valence response value = 0.52SubjVal + 1.22

power response value = -0.29SubjPow - 4.95

Figure 6.1: Improved linear immediate response functions based on the student's emotion in the immediately preceding utterance.

6.4 Procedure

Subjects were asked to interact with Gracie and fill out several questionnaires. After completing the consent form, subjects were briefed on the experiment. They were told that the 20 minute experiment was a study in communicative technology that would investigate

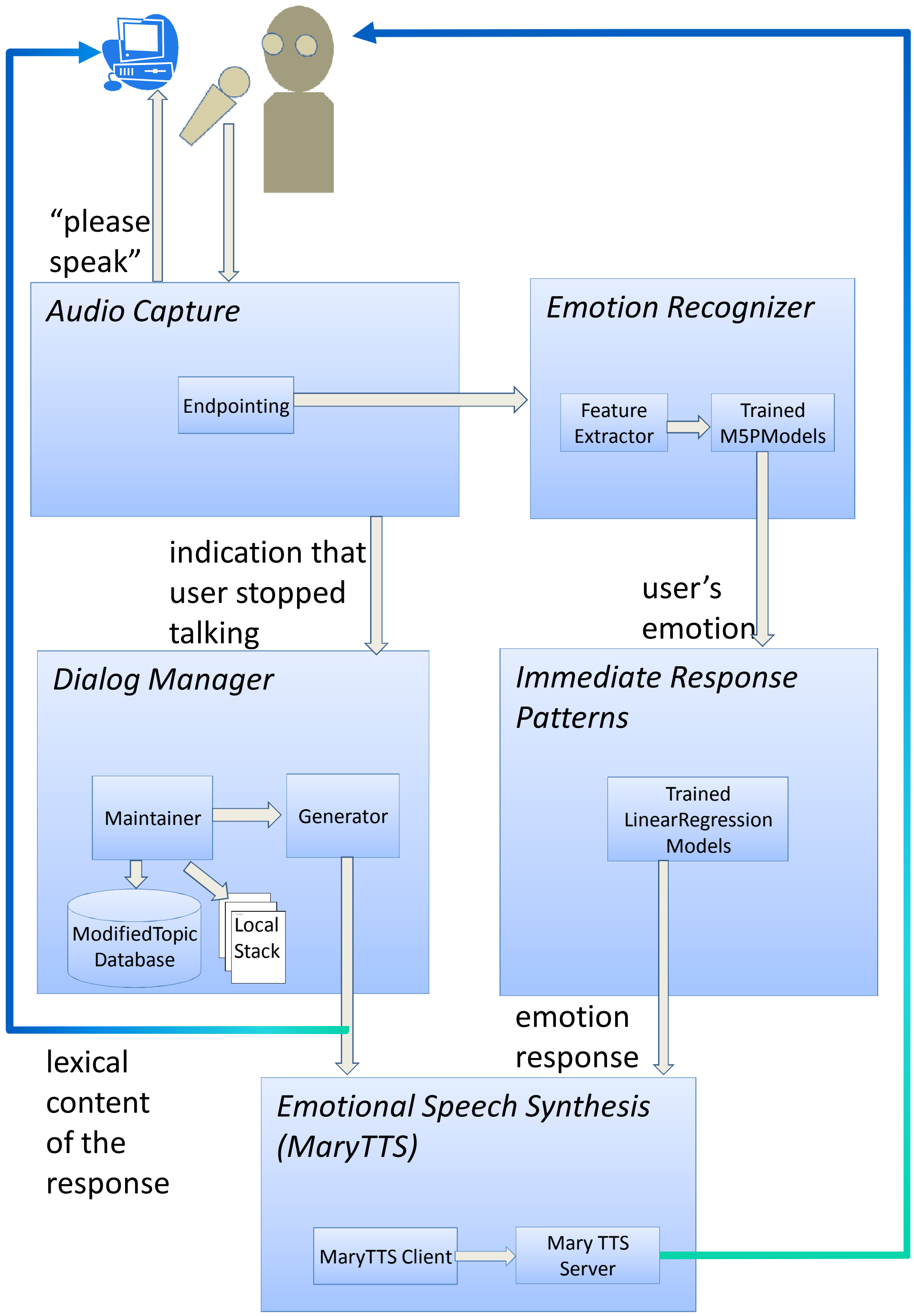

Figure 6.2: Architecture of Gracie that was used for the Experiment

the effectiveness of a dialog system at informing about the graduate school option. Afterwards, subjects filled out a demographic sheet. The purpose of the demographic sheet was to enable analysis of whether people with different gender, age, occupation, and linguistic backgrounds have different perceptions of Gracie. The next step was the interaction with Gracie.

Subjects were given a thirty second demonstration interaction with Gracie. The experimenter indicated that when the words “Please Speak” appeared, they could talk. Otherwise, they would be ignored. Also, during the interaction, the subjects were told that the on-screen content contained the words that Gracie was saying, presented visually also to avoid issues with pronunciation clarity.

Before the users started interacting with Gracie, the experimenter ran a script called *beforeRunningSubject*. This script prepared the non-contingent configuration for the current subject by copying the sequences of calculated emotion responses from the previous user’s rule-based interaction.

As mentioned above, there were 3 possible configurations (rule-based, non-contingent, neutral), and there were 3 fixed content sequences. The conditions were balanced throughout the study. There were a total of 36 possible orderings of configurations and content sequences. To abstract the details from the experiment and to reduce the chance for experimenter error, a script file was written for each interaction. The experimenter ran the script corresponding with the user (a number) and the current run (a letter). For example, if the first subject is interacting the Gracie for the third time, the experimenter will run the script named *1c*. Each script contained flags for the configuration (rule-based, non-contingent, or neutral) and the Topic Database to use (contentA, contentB, and contentC). Next, the experimenter would instruct users to interact with Gracie three times.

After each interaction, users were asked to evaluate their interactions with each system by filling out a questionnaire. Users were asked to base their answers strictly on Gracie’s voice, not on the content presented. The questionnaire was largely based on work from Gratch et al. [17]. The following are the questions included in the questionnaire.

1. I felt I had a connection with the coordinator.
2. I think the coordinator and I understood each other.
3. The coordinator seemed willing to help.
4. The coordinator seemed trustworthy.
5. The coordinator seemed likeable.
6. My conversation with the coordinator seemed natural.
7. I enjoyed the interaction with the coordinator.
8. The coordinator was human-like.
9. The coordinator was persuasive.
10. I would recommend the coordinator to others.

The first two questions are aimed at measuring feelings of emotional rapport and cognitive rapport respectively. Questions 3 - 6 measure qualities that are generally important for dialog systems and embodied conversational agents, but not as important for this research. Questions 7 - 10 were not included in Gratch et al.'s work, but were included to measure other qualities of the interaction that can be important for dialog systems.

After finishing the three interactions, users were asked to rank their preference for each system (best to worst) on a final questionnaire. Both questionnaires are in Appendix F.

6.5 Subject Pool

A total of 36 subjects participated in the study (23 males, 13 females). The subjects in the study were all students from the University. All but three subjects reported speaking more than one language. The students were recruited in two ways. First, some subjects were students enrolled in the Introductory Computer Science class. One requirement of this class is for students to complete two research credits, where students could get a research credit by interviewing a graduate student and writing a report, attending a seminar and taking notes, or by participating in an experiment such as this one. Subjects were also recruited by approaching students in the Computer Science main lab, or in the Student Union; they were offered a $10.00 incentive for participation. In total 12 subjects received research credit for their course and 24 participants were paid the $10.00. None of the subjects had interacted with Gracie before the experiment.

Chapter 7

Results

The results confirm the hypothesis of this work, that a spoken dialog system with emotional intelligence is better at gaining rapport with users. The results also indicate that users preferred the system that adapts to emotional state best. This chapter presents in detail these and other results from the user study along with discussion.

7.1 Rapport and Measures of Interaction Quality

After each of the three interactions with Gracie (neutral, non-contingent, rule-based), subjects filled out a questionnaire with ten questions. These questions are included in the previous chapter. The ratings for the questions were first compared using paired samples t-test to determine significant differences in ratings between pairs of systems. Table 7.1 shows the average ratings and standard deviations for each question for the three versions of Gracie. Differences are reported as significant when $p < 0.05$.

For all questions, the mean ratings for the rule-based system are higher than the those for both controls. Regarding feelings of emotional rapport (Question 1) and cognitive rapport (Question 2), the rule-based system was rated higher than both the neutral and non-contingent versions. These results match the results of Gratch et al. [17] (although their results were not significant). This could mean that adapting to users, whether in voice or by physical gesture, is an essential element for gaining rapport with users.

Regarding Questions 3 and 4 (helpfulness and trustworthiness), the rule-based system was rated highest, followed by the non-contingent version, but these differences were not significant. The ratings for Question 4 had the smallest difference between the three sys-

Table 7.1: Subjects' ratings of the three versions of Gracie.
(*) - significantly higher than neutral
(+) - significantly higher than non-contingent

Question		**Rule-Based**	**Non-Cont.**	**Neutral**
1 - Connection	mean	*+**4.61**	**3.67**	**3.78**
	std.dev.	1.38	1.71	1.62
2 - Mutual understanding	mean	*+**4.72**	**3.94**	**3.75**
	std.dev.	1.61	1.45	1.92
3 - Helpful	mean	**4.58**	**4.33**	**4.11**
	std.dev.	1.75	1.67	1.75
4 - Trustworthy	mean	**4.28**	**4.22**	**3.78**
	std.dev.	1.65	1.87	1.77
5 - Likeable	mean	***4.72**	**4.19**	**3.56**
	std.dev.	1.70	1.79	1.93
6 - Natural conversation	mean	**3.94**	**3.58**	**3.19**
	std.dev.	1.96	1.83	1.83
7 - Enjoyment	mean	***4.69**	**3.89**	**3.47**
	std.dev.	1.88	1.83	1.95
8 - Human-like	mean	***3.94**	**3.31**	**2.78**
	std.dev.	1.82	1.88	1.88
9 - Persuasive	mean	***4.17**	***3.67**	**2.97**
	std.dev.	1.89	1.82	1.70
10 - Recommendable	mean	**4.28**	**3.89**	**3.64**
	std.dev.	1.70	1.98	1.88

tems. This could possibly be because the interactions lacked smalltalk [8], which was shown to improve trustworthiness of a virtual real-estate agent. Users rated the rule-based system as the most likeable (Question 5) among the three systems. Gratch et al.'s results are contrary to this; they found that users rated the non-contingent version more likeable than the responsive system. It may be the case that adapting the emotional coloring of utterances contributes more than physical gestures to likeability. Question 6 (feeling a natural interaction with the coordinator) had low ratings overall, including for the rule-based system. This could be due to the fact that Gracie had fixed responses, with unnatural turn-taking (based on user silence). Something similar was seen in Gratch et al.'s results, where naturalness was again rated lower than most of the other qualities. Perhaps this signifies that these agents, although better, are still lacking naturalness.

Regarding the enjoyment of the interaction (Question 7), subjects enjoyed their interaction with the rule-based system significantly more than with the neutral system, but not significantly more than with the non-contingent version.

Users felt that the rule-based and non-contingent system were significantly more human-like than the neutral version (Question 8), although overall this question had the lowest ratings in all conditions. There was no significant difference between the rule-based and non-contingent versions. This may be because users felt that any emotional coloring is more human-like than monotonous speech.

Subjects felt that both of the systems with emotion were significantly more persuasive than the neutral (Question 9). The rule-based system was rated highest of the three, but there was no significant difference between the rule-based and non-contingent. Lastly, subjects did not significantly prefer to recommend any version of Gracie to others (Question 10).

7.2 Overall Preference

After subjects finished interacting with all three systems, were asked to rank their overall preference for the systems. Out of 36 people, 20 preferred the rule-based, 9 preferred the non-contingent, and 7 preferred the neutral version. The ratings were tested for significance using the χ^2 test. The subjects significantly preferred the rule-based version of Gracie.

7.3 User Comments

Comments from users regarding the rule-based system include, "[the rule-based system] was more like a conversation rather than just listening." Other users noted, "It seemed to be talking to a real person," and "[the rule-based version] seemed easiest to connect to and stay engaged," and "I was very comfortable with the [rule-based] system." Some users still felt that the rule-based and non-contingent versions spoke too fast at some points. This problem seems to exist due to diphone synthesis; some words are not pronounced correctly unless they are pronounced at neutral speed.

Overall, the results indicate that it is important to not only add emotion to spoken dialog systems, but also to adapt emotions to user states. Users overall rated the non-contingent version of Gracie as second-best to the rule-based version. One subject felt the interaction was too extreme because the coordinator always sounded happy: "It was very unreal like she made her sentences sound really happy." An example of the importance of adapting to user emotions can be seen by the comments of two consecutive subjects. The first subject praised the rule-based system saying, "The system was able to adapt to the speed at which it was responded to...". The following user said, "It was boring ..." when the same series of emotional colorings were presented in the non-contingent version.

Negative comments about the neutral system were mainly focussed on its lack of variation in prosody. One comment stated, "[The neutral version] was very monotonous..."

7.4 Individual Differences

Surprisingly, seven subjects preferred the neutral version of Gracie. Based on the comments, one reason for this may be its limited variation of prosody, clearer pronunciation, and slower pace. One user who preferred the neutral version stated, "[The neutral version] was easier to understand." This could be due to the personalities of the subjects as seen in previous work [38, 8].

Another reason may be due to the difficulty of rating the systems based on voice only. Although users were asked three times to base their answers solely on the sound of the coordinator's voice, not on the content, some still based their answers on the content. For example, one user who preferred the neutral version noted, "Much better conversation. Seems like it understood more of my answers, but only choose a keyword from my answer sentence." Another said, "The coordinator made good points about going to graduate school... [I] was able to understand the questions that were asked."

Looking at the linguistic background of the participants, results show that 13 out of the 19 subjects who listed English as their dominant language preferred the rule-based system best. Only 7 of the 17 that listed a different dominant language ranked the rule-based system best. Also, 8 out of the 17 subjects that listed a different dominant language ranked the non-contingent version best (only 1 English dominant speaker preferred the non-contingent version). One possible explanation for this could be that non-native English speakers may speak with less prosody in voice, which would make Gracie produce near-neutral responses, but they actually prefer the emotional coloring.

The full set of comments are included in Appendix D.

Chapter 8

Future Work

This dissertation has shown that incorporating emotional intelligence into a spoken dialog system can increase its effectiveness for building rapport. This work is a first step in creating dialog systems that adapt to users, not only in the traditional sense of content choice, but also in choice of emotional coloring for the content. This work has much room for improvement and extension.

8.1 Possible Improvements to the Current Model

The emotional responses described in this dissertation are in no way a final solution, there are still many ways to improve the quality. In the persuasive dialog corpus, it is almost certainly the case that the emotional responses were not based only on the immediate context. The coordinator may have been adjusting her responses depending on aspects of the students, such as personality, gender, age, and social status (freshman, sophomore, junior, or senior). The coordinator's responses may also have been based on her cumulative interpretation of the user's state. For example, Lee [22], shows that previous emotions during the interaction may influence future emotional states of the interlocutors. It is also likely that the emotional coloring of her responses was influenced by the student's lexical content.

Regarding utterance boundaries, it would be worthwhile to explore a finer grained model to complement the utterance-by-utterance response strategies modeled in this work. Some utterance units had drastic acoustic variation, for example, the coordinator started slow and soft, then immediately changed to fast and loud. Future work may relate emotional interplay using fixed time periods, instead of turns, to find more immediate response patterns.

Embodied conversational agents have attempted to gain rapport in the past by responding to user physical gestures. The methods in this dissertation can apply to and possibly improve these systems. During the user study, many times subjects would perform rich facial and bodily gestures.

Lastly, regarding the speech synthesis, many users commented on the voice of the synthesizer, mentioning that it sounded too much like a machine, especially with high or low speaking rates. One option may be to use recorded voice.

8.2 Broader Impact

Building upon earlier work in adaptive dialog systems, which has shown that the choices among a few alternatives for lexical content and prosodic tailoring of simple acknowledgments are liked and are useful, I have demonstrated the utility of emotional adaptation using generic emotion dimensions. Even with a crude model using only immediate response patterns, the benefit of properly controlling the nonverbal aspects of emotional responsiveness was clear. Specifically, this work has provided a set of rules for appropriately varying prosody to show appropriate emotional responses based on user emotional state. These rules were built and evaluated in Gracie, the first dialog system with emotional intelligence, and shown to improve its ability to gain rapport with users.

References

[1] http://cafe.bevocal.com/.

[2] http://sourceforge.net/projects/opensmile/.

[3] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. Desperately Seeking Emotions: Actors, Wizards, and Human Beings. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. ISCA, 2000.

[4] J. K. Burgoon, L. A. Stern, and L. Dillman. *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge University Press, 1995.

[5] F. Burkhardt. *Simulation of emotional manner of speech using speech synthesis techniques*. PhD thesis, Technische Universität Berlin, 2000.

[6] F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber. An emotion-aware voice portal. *Proc. Electronic Speech Signal Processing (ESSP)*, pages 123–131, 2005.

[7] J. E. Cahn. Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology, 1989.

[8] J. Cassell and T. Bickmore. Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and User-Adapted Interaction*, 13(1):89–132, 2003.

[9] T. L. Chartrand and J. A. Bargh. The Chameleon Effect: The Perception-Behavior Link and Social Interaction. *Journal of Personality and Social Psychology*, 76:893–910, 1999.

[10] M. H. Cohen, J. P. Giangola, and J. Balogh. *Voice user interface design*. Addison-Wesley Longman, 2004.

[11] S. K. D'Mello, S. D. Craig, A. Witherspoon, B. McDaniel, and A. Graesser. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1):45–80, 2008.

[12] P. Ekman. An Argument for Basic Emotions. *Emotion: Themes in the Philosophy of the Mind*, 2005.

[13] K. Forbes-Riley and D. Litman. Predicting emotion in spoken dialogue from multiple knowledge sources. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, 2004.

[14] K. Forbes-Riley and D. Litman. Adapting to student uncertainty improves tutoring dialogues. In *Proceedings of the International Conference on Artificial Intelligence in Education*, 2009.

[15] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I.H. Witten. Using model trees for classification. *Machine Learning*, 32(1):63–76, 1998.

[16] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L. P. Morency. Virtual Rapport. *Proceedings of the 5th International Conference on Interactive Virtual Agents (IVA)*, 2006.

[17] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In *Intelligent Virtual Agents: 7th International Conference*, page 125. Springer-Verlag New York Inc, 2007.

[18] J. Gratch, N. Wang, A. Okhmatovskaia, F. Lamothe, M. Morales, R. J. van der Werf, and L. Morency. Can Virtual Humans Be More Engaging Than Real Ones? *Lecture Notes in Computer Science*, 4552:286–297, 2007.

[19] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky. PocketSphinx: A free, real-time continuous speech recognition system for

hand-held devices. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[20] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura. A Speech Synthesis System with Emotion for Assisting Communication. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. ISCA, 2000.

[21] J. Klein, Y. Moon, and R. W. Picard. This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14(2):119–140, 2002.

[22] C. C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan. Emotion recognition using a hierarchical binary decision tree approach. In *Interspeech 2009*, pages 320–323, Brighton, UK, September 2009.

[23] I. Mazzotta, F. de Rosis, and V. Carofiglio. Portia: A User-Adapted Persuasion System in the Healthy-Eating Domain. *IEEE Intelligent Systems*, pages 42–51, 2007.

[24] N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *Acoustics, Speech and Signal Processing, Proceedings of the 1998 IEEE International Conference*, volume 2, 1998.

[25] V. Petrushin. Emotion in Speech: Recognition and Application to Call Centers. *Artificial Neural Networks In Engineering (ANNIE)*, pages 7–10, 1999.

[26] H. Pon-Barry and S. Shieber. Identifying uncertain words within an utterance via prosodic features. In *Interspeech*, 2009.

[27] H. Prendinger and M. Ishizuka. The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence*, 19(3-4):267–285, 2005.

[28] B. Reeves and C. Nass. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, 1996.

[29] A. G. Rivera and N. G. Ward. Three prosodic features that cue back-channel feedback in Northern Mexican Spanish. *Proceedings of the Seventh Annual High Desert Linguistics Society Conference. University of New Mexico*, pages 19–26, 2008.

[30] J. A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

[31] M. Schröder. Emotional Speech Synthesis: A Review. In *Seventh European Conference on Speech Communication and Technology*. ISCA, 2001.

[32] M. Schröder. Dimensional Emotion Representation as a Basis for Speech Synthesis with Non-extreme Emotions. In *Affective dialogue systems: tutorial and research workshop*, pages 209–220. Springer-Verlag, 2004.

[33] M. Schröder. *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis.* PhD thesis, Research Report of the Institute of Phonetics, Saarland University, 2004.

[34] M. Schröder and J. Trouvain. The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377, 2003.

[35] B. Schuller, S. Steidl, and A. Batliner. The Interspeech 2009 emotion challenge. *Proceedings Interspeech, Brighton, UK*, 2009.

[36] C. A. Shepard, H. Giles, and B. A. Le Poire. Communication accommodation theory. *The new handbook of language and social psychology*, pages 33–56, 2001.

[37] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207, 2000.

[38] N. Ward and W. Tsukahara. A study in responsiveness in spoken dialog. *International Journal of Human-Computer Studies*, 59:603–630, 2003.

[39] N. G. Ward and Y. Al Bayyari. A case study in the identification of prosodic cues to turn-taking: Back-channeling in Arabic. In *Interspeech*, 2006.

[40] N. G. Ward and R. Escalante-Ruiz. Using subtle prosodic variation to acknowledge the user's current state. In *Interspeech*, 2009.

[41] I. H. Witten and E. Frank. Data mining: practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record*, 31(1):76–77, 2002.

[42] S. Young, G. Evermann, D. Kershaw, G. Moore, D. Ollason, V. Valtchev, and P. Woodland. *The HTK book*. Cambridge University, 1995.

Appendix A

Topic and Strategy Database

This appendix contains the structure of the database used for the VoiceXML prototype. Sections contain several topics. Some topics ask the user questions. Based on the user's response, other sections are enabled and disabled. In addition, some topics contain acknowledgements. These are short responses that are spoken to the user immediately after their response. This is described more in Chapter 3. In the following tables, IE stands for Interested Enables, identifying the sections that are enabled if a user says "yes." ID stands for Interested Disables. This contains the sections that are disabled if the user says "yes." NIE stands for Not Interested Disables. This contains the sections that are enabled if the user says "no." Lastly, NID stands for Not Interested Disables. This contains the sections that are disabled if the user says "no."

Section 1 Essential Information			
a. systemGiveGreeting	*Hi, I'm glad you called. My name is Sophia.What is your name?*		
Acknowledgement:	*That's a nice name.*		
b. systemGivePraiseForComing	*I'm really glad you called because now I can tell you more about graduate school.*		
c. systemGetPreferenceMajorCS	*Is your major Computer Science?*		
IE:2,3,6,8,9,10,18	ID:	NIE:6	NID:
Acknowledgement:	*Ok, that's interesting.*		
d. systemGetConcernCSDifficulty	*Are you finding your classes to be hard?*		
IE:	ID:	NIE:4,5,13,15	NID:
Acknowledgement:	*OK*		
e. systemGetInfoClassification	*Are you a freshman, sophomore, junior, or senior?*		
freshman enables:10	freshman disables:3,4,5	other enables:	other disables:
sophomore enables:10,12	sophomore disables:4,15	other enables:	other disables:
junior enables:5,14	junior disables:8,18,19	other enables:	other disables:
senior enables:4,5,7,17	senior disables:2	other enables:	other disables:
Acknowledgement:	*Ok, that's interesting.*		
Section 2 Encouragement			
a. systemGiveAdviceForFutureImportanceOfInfo	*You know, this advice may be very relevant to you.*		
b. systemGivePersMajorAnecdote	*I made a mistake when I started. I had to change my major three times because I couldn't decide what I wanted to do.*		
c. systemGiveAdviceForFuture	*We've got a lot of our Masters students take jobs as performance analysts and consultants with really good companies. So if you see yourself in that kind of job in that kind of environment, then you can already make a mental note in your head that, "I really need to start thinking about probably exploring all the graduate school options."*		
d. systemGiveAdviceForFutureGrades	*Another thing is that as far as getting into graduate school is concerned there are a couple of things you may want to start thinking about even now. One is that how you do in your classes matters very much.*		
Section 3 General Information about the Graduate Program			
a. systemGetGeneralProgramInterest	*Do you want to hear more about our graduate program? It won't take long.*		
IE:7,15	ID:	NIE:11,19	NID:3,4,5
	Alright.		
b. systemGiveInfoProgramsOffered	*I know in this department, we have a Masters in Computer Science (a Master of Science), and we also offer a PhD.*		
c. systemGiveInfoForProgram	*The difference between the Masters and the PhD is not just, you know, what you do once you get out of the program.*		
d. systemGiveInfoForProgramMasterTime	*If you're in a Masters program, it's probably going to take you two years to finish; maybe a little bit more if you go slower.*		
e. systemGiveInfoForProgramPhDTime	*I guess it's obvious that the PhD is longer. So we're talking about a difference of about maybe two years for your Masters, and probably four or five years, probably even as long as six for your PhD.*		

Section 4 Statement of Purpose Requirements			
a. systemGetSOPRequirementsInterest	*Would you like me to tell you some tips on how to write a great statement*		
IE:15	ID:	NIE:	NID:4
Acknowledgement:	*Sounds good.*		
b. systemGiveInfoGradSchoolRequirementsSOP	*Probably one of the most important things to getting into any graduate program, is writing a good statement of purpose. Some people call them statement of purpose; I've seen them called statement of intention, or a personal statement.*		
c. systemGiveInfoForSOPSections	*It all boils down to a short essay in which you talk about your goals, your background, maybe what your interests are, and why you want to pursue the advanced degree with that department.*		
d. systemGiveInfoForSOPHowToWrite	*The statement of purpose should be what you want to sell. It should be written in a way that you want to sell yourself to those people that are reading it. They should be able to read it and want you in their program. You're not going to be in the room to talk to them and say, "hey look i'm good you should let me in." It should be able to say that for you even though you're not there.*		
Section 5 Other Graduate School Requirements			
a. systemGetOtherRequirementsInterest	*There are just a few requirements for the graduate school, would you like to hear about them?*		
IE:4,15	ID:	NIE:	NID:5
Acknowledgement:	*Alright then.*		
b. systemGiveInfoGradSchoolRequirementsGPA	*I think our University's graduate school asks that you have at least a three point oh, in order to get into the graduate school.*		
c. systemGiveInfoGradSchoolRequirementsExp	*Even as an undergraduate, you can go to some research group meetings and talk to other students who are doing research. You know, having some idea of what that's all about. That will give you an advantage into getting into graduate school and, you know, putting your best foot forward as a*		
d. systemGiveInfoGradSchoolRequirementsGRESOP	*You'll also have to take a GRE test and write a statement of purpose.*		
e. systemGiveInfoGradSchoolRequirementsConc	*So those are some of the requirements for graduate school. It may seem like a lot, but it is not that bad at all.*		
f. systemGiveInfoRequirementsGPAOtherSchools	*You'll find that it's the same note, even if you decided, "hey I want to go to graduate school somewhere else." These kind of things are pretty much the same. They'll look at your your grades to kind of see where you are. And usually there is a placement test like a GRE or something.*		
Section 6 Farewell			
a. systemGiveGoodbye	*It was nice talking to you, thank you for calling.*		

<table>
<tr><th colspan="4">Section 7 Graduate Program Flexibility</th></tr>
<tr><td>a. systemGetProgramFlexibleInterest</td><td colspan="3">One thing I like about our graduate program is that it is very flexible. Can I tell you about it?</td></tr>
<tr><td>IE:13,16</td><td>ID:</td><td>NIE:</td><td>NID:7</td></tr>
<tr><td>Acknowledgement:</td><td colspan="3">OK then.</td></tr>
<tr><td>b. systemGivePersFlexibleDirectPhD</td><td colspan="3">Our program here is very flexible because you can either go from a Bachelors degree to a PhD or from a Masters to a PhD.</td></tr>
<tr><td>c. systemGivePersMasterToPhDTransition</td><td colspan="3">In our program, you can easily decide to either change to the PhD program, or you can finish your Masters and then start the PhD program.</td></tr>
<tr><td>d. systemGivePersFlexibleMasterToPhdTransition</td><td colspan="3">If you start in the Masters program and you say, " I really like this, This is really interesting and I know now, for sure, I want to get my PhD", all those classes will just smoothly transfer into a PhD program of study. So you can't make a wrong choice there.</td></tr>
<tr><td>e. systemGivePersFlexiblePhDToMasterTransition</td><td colspan="3">If you apply to the PhD program, and you get admitted, and you're in the program, and say maybe a year in; you decide "you know what?" for whatever reason maybe this isn't for me, then, all those classes that you've already taken can go towards your Masters.</td></tr>
<tr><th colspan="4">Section 8 Careers After Graduate School</th></tr>
<tr><td>a. systemGetCareerReasonsInterest</td><td colspan="3">Would you like to hear about the great career opportunities that are available when you finish the graduate program?</td></tr>
<tr><td>IE:9,13</td><td>ID:</td><td>NIE:10</td><td>NID:8,18,19</td></tr>
<tr><td>Acknowledgement:</td><td colspan="3">Alright.</td></tr>
<tr><td>b. systemGiveInfoPostGraduationChoices</td><td colspan="3">After you graduate you can either work right away, or you can continue into graduate school.</td></tr>
<tr><td>c. systemGivePersExamplesOfGradStudentJobs</td><td colspan="3">We've got a lot of our Masters students take jobs as performance analysts, and consultants, and jobs like that with really good companies. I can tell you all the companies that I see our graduates going to. Exxon, Intel, IBM , they recruit a lot of our students because they know that they are good and they know what they're doing.</td></tr>
<tr><td>d. systemGivePersGradSchoolDifferentDegreePlan</td><td colspan="3">One thing that makes graduate school different is that there's a huge research component. If you're getting your Masters you take classes just like you would as an undergraduate. But at the end of it, you write a thesis. In order to write that thesis you have to pick a question. You have to go out and do research on how to answer that question.</td></tr>
<tr><td>e. systemGivePersBenefitsOfResCmpyDyn</td><td colspan="3">If you're doing research I don't think you can get bored. You decide what question you want to answer, and then you go out and answer it.</td></tr>
<tr><td>f. systemGivePersBenefitsOfResearcherCompany</td><td colspan="3">Another thing is that if your thesis work is for a company, you know, research and development; of course they're probably going to be invitations after you graduate.</td></tr>
<tr><td>g. systemGivePersBenefitsOfResearcherDynamic</td><td colspan="3">Getting a graduate degree is definitely a way to overcome the boredom of sitting behind a desk programming all day long for the rest of your life.</td></tr>
</table>

Section 9 General Funding for Graduate School

a. systemGetGradFundingInterest	*There are a lot of funding opportunities for graduate students. Want to hear more?*		
IE:11,19	ID:	NIE:	NID:9
Acknowledgement:	*Alright.*		
b. systemGivePersDiffInProgramPhDFunding	*If you're in the Masters program, you're kind of on your own in terms of funding. You've got to go out there and find a job on campus, or apply to be a TA. But if you go in to the PhD program; because it's so much of a longer committment, the department actually matches you up with funding. So you won't actually have to go out there and, say, beat the pavement, trying to find something. And we'll tell you, "hey there's a TA slot we're gonna put you into that TA slot." You'll have an advisor whose gonna say, "I need a Research Assistant; come on and be my Research Assistant." So there is that subtle but very important difference of getting paid while you're in*		
c. systemGivePersForGradFunding	*Another thing about graduate school that you should know. There are a lot of different ways to pay for it over and above; you know; taking money out of your pocket. There are scholarships. There are fellowships. There are assistantships. There are campus jobs, where people prefer to hire a graduate student because the assumption is that the student is more mature; that they have a greater body of knowledge to draw from. So they'll be able to help them better. So the outlook for paying for graduate school looks a lot different from paying for your undergraduate degree. There's so many more chances to get funded. It's actually a really good*		

Section 10 Returning in the Future

a. systemGetFutureVisitsInterest	*I'm always available, I can help you a lot in your future. Would you like to hear about scholarships or fellowships?*		
IE:11	ID:	NIE:	NID:10
Acknowledgement:	*OK.*		
b. systemGivePersForFutureVisits	*As a matter of fact, I have some scholarships that are accepting applications right now. Come talk to me in person to hear more.*		

Section 11 Scholarship Information

a. systemGivePersForFutureVisitsApplicationHelp	*Once you get ready to apply I can walk you through that whole application process, tell you exactly what you need to do I can proofread your statement for you.*
b. systemGivePersApplicationHelpScholarships	*You'll probably come back and see me in two years. I hope you do when you're getting ready to graduate. We can talk about; well; there's actually a couple of things that I'll be able to do for you then. I'll be able to, if you haven't found a job yet, point you in the right direction. I know a lot of students know that we have a career services center, but they don't know all the stuff that they do. The cool thing is that I actually have a couple of friends that work in that office. They have been really helpful in finding our students jobs. So, i'll be able to help you with that.*

Section 12 Scholarship Information for Sophomores			
a. systemGivePersScholarshipSophomore	*You know, there are a lot of scholarships available, especially at this time in your degree. If you have good grades, you have a great chance at getting school paid. This also looks good a graduate school application, if you received scholarships.*		
Section 13 Notable Qualities of the Graduate Program			
a. systemGetProgramQualitiesInterest	*Our graduate program has very great qualities, do you want to hear about them?*		
IE:9	ID:	NIE:	NID:13
Acknowledgement:	*Sounds good.*		
b. systemGivePersPhDTeach	*Usually when we get PhD students in, they start off in TA positions. If they're really good, and the department has a need, they can apply to be an instructor. If there's a match they can get hired. To my knowledge three PhD students in the last three years have done that.*		
c. systemGivePersQualitiesProgramPhDSummary	*So there are a lot of great things that are available for your higher degree from our program.*		
Section 14 Junior Classification Praise			
a. systemGivePraiseClassificationJunior	*A rising junior. Well, that's good. So in maybe a couple years you would have probably come by to see me anyway because that's when people really start thinking about, "oh my gosh, I'm getting ready to graduate what am I gonna do now?"*		
Section 15 Graduate Record Exam (GRE) Requirements			
a. systemGetGREDetailsInterest	*Would you like to hear a little bit more about the GRE?*		
IE:5	ID:	NIE:	NID:15
Acknowledgement:	*Excellent.*		
b. systemGiveInfoForGRESubjects	*So, the GRE test has three areas. The math area, a verbal one, and a writing one where you write a short essay. It's really very similar to the SAT.*		
c. systemGiveInfoGradSchoolReqGREGrades	*The graduate school will look at your GRE score in your application.*		
d. systemGiveInfoGREOtherSchools	*No matter where you're going to apply to graduate school, whether it's here or somewhere else, usually people require that you take an exam. It's not a real tough exam, but you still; you know; you need to mentally prepare for it. Probably spend a month or so working examples so that you'll get used to the format and the questions and stuff like that. Then, you can do well. It's another thing that's going to help the admissions people decide if you are; you know; if you look like you're going to be a good match for graduate school.*		

Section 16 Masters to PhD Encouragement

a. systemGetMSToPhDEncourageInterest	*Can I tell you about going from a Masters to a PhD?*		
IE:	ID:	NIE:	NID:16
Acknowledgement:	*Alright.*		
b. systemGivePersQualitiesProgramPhD	*This is a great school to get your PhD. If you stay and you stick with it and you get your PhD, you're at the top of the food chain academically. I mean it's the highest degree you can get. All the options are open to you. I just thought of something else. Money. You get paid a lot more. I looked at some numbers. I think last week I saw the latest numbers. The difference between someone coming out with a Masters and the national average is 65000 dollars a year; that's what you make. So, sometimes it's lower sometimes it's higher but just around 65. Coming out with a PhD it's 95 so that's a huge difference.*		

Section 17 Senior Classification Praise

a. systemGivePraiseClassificationSenior	*A senior, you are almost there. Good job.*

Section 18 Teaching in a University

a. systemGetInfoTeachingInterest	*Are you interested in ever teaching in a University?*		
IE: 19	ID:	NIE:	NID:18
Acknowledgement:	*OK.*		
b. systemGivePersUniveProfRequirement	*University professors typically have a PhD.*		
c. systemGivePersUnivProfRequirement	*If teaching is your long term goal, well, you can teach at a community college with your Masters. And if you want to teach at a University, then you need your PhD.*		

Section 19 Reasons to Go to Graduate School

a. systemGetInfoPersueGradDegreeInterest	*Do you want to hear some reasons why students persue a graduate degree?*		
IE:	ID:	NIE:	NID:19
Acknowledgement:	*Alright.*		
b. systemGivePersReasonsForGraduateSchool	*Well, there are a couple reasons why people go on to get a graduate*		
c. systemGivePersReasonForGradSchoolCareerGoal	*A great reason to go on in your studies is because you have a goal; a career goal, perhaps, and you know that getting an advanced degree, a Masters or a PhD is going to help you achieve that goal. Or, in some cases, that you absolutely need one to get where you want to be in life.*		
d. systemGivePersReasonForGradSchoolNoJob	*One of the most common reasons for continuing to graduate school is that students graduated and they could not find a job. Hopefully that's not gonna happen in your case.*		
e. systemGivePersResearcherRequirement	*Also, if you want to be a high-level researcher for a company, you can also do that with a PhD.*		
f. systemGivePersReasonForGradSchoolConclusions	*So those are some of the reasons why people start thinking about graduate school. It always always always has to tie into your goals. I think that is what really sustains your interests and keeps it; and keeps you going throughout the length of the program.*		

Appendix B

Experimenter Steps for the VoiceXML Prototype User Study

This appendix contains the experimenter steps that were followed during the experiment with the VoiceXML Prototype.

Evaluation of a Prototype System for Informing Students about the Graduate School Option by Dialog:

Experimenter Steps

Date _________

Preparation

1. Setup main components of system
    a. Login with developer account
    b. Voice server redirection to correct web server
    c. Web server is running and is reachable from voice server
    d. MySql database is running
    e. Ensure database tables are reset (this can be done by running the system once)
2. Welcome the subject and thank them for participating
3. Tell them the experiment will last about 40 minutes
4. Overview what they'll do:
    a. fill out some paperwork
    b. learn about the automated advisor
    c. use the system
    d. fill out a questionnaire
    e. hear about the research aims
5. *Subject fills out consent form*
6. Sign as witness
7. *Subject fills out demographic information sheet*
8. Assign the subject number (id); write it above, and on the consent, demographic sheet and the questionnaire
9. Record the time ____
10. Briefly explain the automated advisor
    a. Graduate information
    b. Persuasive intent
    c. Prototype system (many limitations)

Experiment

11. Explain how to interact with the system
    a. Dialing the number
    b. Entering the pin
    c. Allowable user responses (grammars)

Closing

12. Give questionnaire
13. After the questionnaire is complete, look it over, and ask them follow-up questions about anything which is unclear or interesting. Write down key points of their responses in the margins.
14. Briefly explain the aims of the research; and answer any questions.
15. Note down any interesting questions or comments that came up.
16. Briefly explain how we'll use their data
    a. we want to see if the system is persuasive.
    b. we want to determine any shortcomings of the system and how we can improve it.
17. Promise to tell their TA that they participated.
18. Thank them warmly.

Appendix C

Questionnaire for the VoiceXML Prototype User Study

This appendix contains the questionnaire given to users after they interacted with the VoiceXML Prototype.

Evaluation of a Prototype System for Informing Students about the Graduate School Option by Dialog:
Questionnaire

Subject ID_________

1. Did the system change your ***attitude*** towards graduate school?
   (Circle a number)

Negative		*No Change*		*Positive*
-2	-1	0	1	2

   Explain:

2. Did you feel more or less ***willing*** to attend graduate school after talking to the advisor? (Circle a number)

Negative		*No Change*		*Positive*
-2	-1	0	1	2

   Explain:

3. What level of ***rapport***, or connectedness did you feel with the advisor?
   (Circle a number)

Negative		*None*		*Positive*
-2	-1	0	1	2

   Explain:

4. How ***trustworthy*** was the advisor?
   (Circle a number)

Negative		*None*		*Positive*
-2	-1	0	1	2

   Explain:

(cont. on back)

5. Please indicate how much the following system properties could have made your ***attitude towards graduate school*** more positive.
   a. **More human-like responsive conversation behaviors**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4
   b. **More Personality**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4
   c. **More feeling in the voice**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4
   d. **Other (please indicate here) ______________**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4

6. Please indicate how much the following system properties could have made your ***willingness to attend graduate school*** more positive.
   a. **More human-like responsive conversation behaviors**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4
   b. **More Personality**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4
   c. **More feeling in the voice**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4
   d. **Other (please indicate here) ______________**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4

7. Please indicate how much the following system properties could have made the advisor ***more trustworthy***.
   a. **More human-like responsive conversation behaviors**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4
   b. **More Personality**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4
   c. **More feeling in the voice**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4
   d. **Other (please indicate here) ______________**
   Not at all *Somewhat* *Extremely*
   0 1 2 3 4

8. Please indicate how much the following system properties could have increased your ***connectedness*** with the advisor.

   a. **More human-like responsive conversation behaviors**

Not at all		*Somewhat*		*Extremely*
0	1	2	3	4

   b. **More Personality**

Not at all		*Somewhat*		*Extremely*
0	1	2	3	4

   c. **More feeling in the voice**

Not at all		*Somewhat*		*Extremely*
0	1	2	3	4

   d. **Other (please indicate here) ______________**

Not at all		*Somewhat*		*Extremely*
0	1	2	3	4

9. Please indicate how much the following system properties could have improved your overall experience with the advisor.

   a. **More human-like responsive conversation behaviors**

Not at all		*Somewhat*		*Extremely*
0	1	2	3	4

   b. **More Personality**

Not at all		*Somewhat*		*Extremely*
0	1	2	3	4

   c. **More feeling in the voice**

Not at all		*Somewhat*		*Extremely*
0	1	2	3	4

   d. **Other (please indicate here) ______________**

Not at all		*Somewhat*		*Extremely*
0	1	2	3	4

(continued on back)

10. Overall in your opinion, which ***emotions or emotional expressions***, if any, do you feel were missing that would make the advisor more ***persuasive, created a feeling of connectedness, and/or more trustworthy***? (please indicate which in your answer)

Explain:

11. Overall in your opinion, what was ***missing***, if anything, in the advisor's responsive behaviors that would make her more ***persuasive, created a feeling of connectedness, and/or more trustworthy***?

Explain:

12. Overall in your opinion, what was ***missing***, if anything, in the advisor's ***personality*** that would have made the system more ***persuasive, created a feeling of connectedness, and/or more trustworthy***? (please indicate which in your answer)

Explain:

13. Overall in your opinion, what was ***missing***, if anything, in the advisor's ***social cues*** (use of words, accent, etc…) that would have made the system more ***persuasive, created a feeling of connectedness, and/or more trustworthy***? (please indicate which in your answer)

Explain:

14. Do you have any other comments or suggestions for improvements in the advisor system?

Appendix D

User Comments from the Gracie User Study

This appendix contains the list of comments by users during the experiment with Gracie. First are the comments about the rule-based system. Next are the comments about the non-contingent and neutral systems. Lastly, comments that include more than one system, and also comments about the experiment in general are included.

The subjects were unaware of which system they were interacting with. The order was balanced. Some comments referenced the systems by the order in which they were presented, for example "the first system felt natural." These references have been replaced by the name, inside brackets, of the actual system that was being used: "the [rule-based] system felt natural."

Subject#	Rule-based System Comments
2	I was very comfortable with the [rule-based] system.
4	On [the rule-based system] I could feel the connection with the operator.
5	Sometimes the coordinator speaks too fast like it's on a hurry. Hard to understand when she does that.
8	Slow down the talking, but keep it cheerful.
12	The coordinator is more interactive, but sometimes it speaks too fast and that make it to hear funny. But, compare with the other 2 it is much better. Also, it seems to be more persuasive and not as boring as the first 2.
14	I can tell when it changes its voice sounds like a computer but I can tell it changes emotion.
18	[The rule-based system] was more like a conversation rather than just listening. Have her sound more enthusiastic.
20	Spoke at strange rate.
22	Sounded like a movie in fast forward, but acceptable.
26	Spoke rather fast in the beginning. I'm not sure if my answers were clear enough.
27	This one talked a bit more funny. Almost like someone who was talking with "helium". This one was more likeable had more responses to my answers but still it was very uncomfortable.
28	The system seemed to adapt to the speed at which it was responded to, but if it could also be able to adapt to the responses at that speed it would be helpful as there were times when it would kind of slur some words together.
29	It seemed to be talking to a real person.
34	At the starting point of the conversation the coordinator's tone was very understandable. Yet there was one point where the coordinator talked really fast. I felt that the coordinator responded to my responses by giving answers. The coordinator made me laugh I was able to connect the conversation did feel human-like.
36	[The rule-based system] seemed easiest to connect to and stay engaged.

Subject#	Non-Contingent System Comments
5	The voice sometimes too slow, like she is sleeping and don't like what she's doing.
9	Coordinator was too fast, but it sounded really excited about graduate school.
11	The coordinator was worse than the other. The things it said, don’t encourage me to go to graduate school.
14	It was more human-like.
15	She talked too fast and too high pitch. It was very unreal like she made her sentences sound really happy. If I didn't have the reading material in front of me, I probably wouldn't have understood a good part of the information. Needs to be more natural tone. Some were too high and some were weird.
20	Seemed to have more of a personality because of what he/she stated.
22	Still it is speaking like a robot, at least that voice stutter a little bit, like someone no familiar with any language.
26	Just a little too fast. I feel as if my answers are not being understood by the coordinator.
29	It was boring, but it has as the others, well as 2, the answer you may be looking for depending in what you say.
30	She understood the subjects I was trying to talk about.
34	Natural, human-like well even though I was able to understand the coordinator, I still felt like I was talking to a computer robot.

Subject#	Neutral System Comments
3	The coordinator seemed happy, making me feel more at ease.
5	The voice sounds funny.
10	I think the machine should have emotion as it speaks.
12	The coordinator was very close to a real conversation 1 or 2 times, but it is very easy to identify that it is not.
15	This one is similar to the [non-contingent] one, but this one talked slower and was easier to understand.
26	Much better conversation. Seems like it understood more of my answers, but only choose a keyword from my answer sentence.
29	Boring sound makes people hung up.
34	The coordinator made good points about going to graduate school. Tone of voice coordinator seemed more human in a way. Was able to understand the questions that were asked.

Subject#	Other Comments
2	I found the systems very informative and enjoying. There was not any difficulty. It's a fun experiment.
6	Maybe the combination of both [non-contingent and rule-based systems] could be an interesting computer program. More human and likable. Good experiment. Not boring. Kind of interesting. Keep on!
9	Make transitions between words smoother.
13	Maybe find another way to speed up speaker or other signs of happiness. [The neutral system] was slower, but felt the intonations more.
14	It was interesting.
16	[The neutral system] was very monotonous and voice sounded as she was tired, [the non-contingent system's] voice is very high and spoke too fast, while [the rule-based system] was almost perfect.
18	The voice used should be a little more fluent and not so instant. I mean like when she stops saying words all of a sudden.
20	Make the conversation more interesting.
21	The voice of all three seems very computer-like, so maybe that can be improved.
23	I think is good enough.
25	Make [the non-contingent system] less slow, make [the rule-based system] more clear.
26	Maybe to ask more questions pertaining to the responses given to the coordinator.
27	I just felt that the way how they asked some questions did not really encourage one to answer it.
28	If you could get it to pause more realistically like how a person breathes in while speaking, that would improve its ability to seem more human-like.
29	Good experiment, could change the way of communicating for most people, at the time they have to make a payment or to get information.
33	Change the voice. It sounds too robot-like, it's scary.

Appendix E

Data from the Gracie User Study

This appendix contains the raw data collected from the user study with Gracie. First the subjects' demographic information and general preference are provided. Next the actual ratings for each question on each system is provided.

						Preference		
Subject#	Paid?	Age	Gender	Occupation	Languages	Best	Mid	Worst
1	y	20-24	m	student	span/engl	non-c.	neu	rules
2	n	60+	m	n/a	engl/span	rules	neu	non-c.
3	n	20-24	f	student	span/engl	neu	non-c.	rules
4	y	20-24	f	student	span/engl	rules	neu	non-c.
5	y	20-24	m	student	span/engl	rules	neu	non-c.
6	n	20-24	m	student	span/engl	non-c.	rules	neu
7	n	25-29	f	student	span/engl	non-c.	rules	neu
8	y	20-24	f	physics student	span/asl/engl/fren	non-c.	rules	neu
9	y	20-24	f	student	span/engl	rules	neu	non-c.
10	n	20-24	m	student	engl/span	rules	neu	non-c.
11	y	30-34	m	Eng.PeerAdvisor	engl/span	rules	neu	non-c.
12	y	25-29	m	student	span/engl	rules	neu	non-c.
13	y	20-24	m	student	engl/span	rules	neu	non-c.
14	y	30-34	f	elec.Eng	engl/span	rules	non-c.	neu
15	n	20-24	f	self.empl	engl/span	rules	neu	non-c.
16	n	20-24	m	student	span/engl/dutch	rules	non-c.	neu
17	n	18-19	m	student	engl/span	neu	non-c.	rules
18	y	20-24	m	CS student	engl/span	rules	neu	non-c.
19	y	25-29	m	student/tech	engl/span	neu	non-c.	rules
20	y	20-24	m	student	span/engl	non-c.	neu	rules
21	y	20-24	m	student	bulg/ger/engl/span	neu	rules	non-c.
22	n	n/a	m	student	span/engl	rules	non-c.	neu
23	y	20-24	m	CS	span/engl	non-c.	neu	rules
24	y	20-24	m	student	span/engl	rules	neu	non-c.
25	y	20-24	m	student	span/engl	non-c.	rules	neu
26	n	25-29	m	tutor	engl/span	neu	non-c.	rules
27	y	20-24	m	student	engl	rules	non-c.	neu
28	y	20-24	m	n/a	engl/span	rules	non-c.	neu
29	n	18-19	f	student	engl	rules	non-c.	neu
30	y	20-24	f	student	span/engl	non-c.	rules	neu
31	n	20-24	f	student	engl	non-c.	rules	neu
32	y	25-29	m	customer.serv.	engl/span	neu	rules	non-c.
33	y	25-29	f	student	engl/span	neu	rules	non-c.
34	y	20-24	f	student	engl/span	rules	neu	non-c.
35	y	20-24	m	student	engl/span	rules	non-c.	neu
36	y	20-24	f	student	engl/span	rules	non-c.	neu

Question 1

Subject #	Neutral	Non-Cont.	Rules
1	4	6	4
2	7	7	7
3	6	3	3
4	6	6	7
5	2	1	4
6	2	4	3
7	5	6	5
8	3	5	3
9	5	3	5
10	2	2	5
11	4	4	7
12	2	1	3
13	5	3	5
14	4	4	4
15	4	1	6
16	5	3	6
17	6	5	3
18	2	1	3
19	6	5	5
20	3	3	2
21	4	4	4
22	3	4	5
23	6	7	5
24	5	3	6
25	5	5	4
26	5	5	3
27	2	3	4
28	1	3	5
29	1	5	7
30	3	5	6
31	4	4	3
32	4	2	5
33	3	2	3
34	3	1	5
35	1	2	5
36	3	4	6

Question 2

Subject#	Neutral	Non-Cont.	Rules
1	4	6	5
2	7	7	7
3	6	4	2
4	6	5	7
5	1	2	4
6	1	4	3
7	4	5	5
8	4	5	6
9	5	3	5
10	2	2	3
11	4	4	4
12	1	2	2
13	5	3	6
14	4	5	6
15	4	3	6
16	5	5	6
17	7	3	4
18	3	1	3
19	5	5	5
20	1	5	1
21	5	4	4
22	2	5	6
23	7	7	4
24	5	4	6
25	4	4	2
26	5	4	3
27	2	3	4
28	1	5	6
29	1	4	7
30	4	5	4
31	5	5	4
32	5	3	6
33	5	4	5
34	3	3	6
35	1	1	6
36	1	2	7

Question 3

Subject#	Neutral	Non-Cont.	Rules
1	5	4	4
2	7	7	7
3	6	5	3
4	7	5	7
5	3	1	5
6	3	5	4
7	6	6	6
8	4	6	5
9	5	4	6
10	4	4	3
11	4	5	3
12	2	1	3
13	6	4	6
14	6	4	6
15	4	4	6
16	3	6	7
17	5	5	1
18	1	2	3
19	7	6	5
20	3	6	1
21	5	3	2
22	3	6	5
23	6	7	3
24	5	5	5
25	5	5	5
26	4	4	2
27	2	4	6
28	1	4	3
29	2	5	7
30	5	6	6
31	4	5	5
32	6	3	4
33	3	1	3
34	3	2	7
35	1	1	5
36	2	5	6

Question 4

Subject#	Neutral	Non-Cont.	Rules
1	5	6	4
2	7	7	7
3	6	4	2
4	6	6	6
5	3	3	6
6	3	5	4
7	5	6	5
8	5	7	6
9	4	2	6
10	3	4	4
11	3	5	4
12	1	1	4
13	6	4	6
14	4	5	3
15	2	2	2
16	4	4	6
17	6	5	1
18	1	1	2
19	4	4	4
20	4	7	2
21	6	4	3
22	3	6	5
23	6	7	6
24	4	3	3
25	6	6	6
26	5	4	4
27	2	3	6
28	1	2	1
29	1	5	7
30	5	6	5
31	4	6	5
32	4	4	3
33	2	1	3
34	2	2	4
35	1	1	4
36	2	4	5

Question 5

Subject#	Neutral	Non-Cont.	Rules
1	5	3	5
2	7	7	7
3	7	5	5
4	6	6	7
5	2	2	7
6	2	5	4
7	5	6	5
8	4	7	6
9	5	3	4
10	2	3	4
11	1	4	6
12	1	1	2
13	6	3	5
14	4	6	4
15	4	4	5
16	5	3	7
17	5	6	1
18	1	1	2
19	4	5	2
20	4	6	4
21	4	3	4
22	2	4	5
23	4	7	3
24	6	5	3
25	6	7	6
26	5	4	5
27	1	2	7
28	1	4	4
29	1	5	7
30	5	5	5
31	1	6	5
32	4	4	3
33	2	1	2
34	2	3	6
35	1	3	6
36	3	2	7

Question 6

Subject#	Neutral	Non-Cont.	Rules
1	3	6	3
2	7	7	7
3	7	4	1
4	7	5	7
5	1	2	4
6	2	4	3
7	4	6	4
8	3	6	4
9	5	2	5
10	2	2	4
11	2	1	4
12	2	2	2
13	5	3	3
14	4	4	2
15	5	1	3
16	3	2	7
17	3	4	1
18	1	1	1
19	3	3	1
20	2	5	2
21	5	4	5
22	2	3	4
23	4	7	4
24	6	2	5
25	3	5	3
26	4	3	4
27	1	3	6
28	1	3	2
29	1	6	7
30	5	6	5
31	1	5	7
32	4	5	1
33	2	1	3
34	2	1	7
35	1	3	6
36	2	2	5

Question 7

Subject#	Neutral	Non-Cont.	Rules
1	6	5	6
2	7	7	7
3	4	4	2
4	4	6	7
5	3	1	3
6	1	5	4
7	5	6	4
8	1	4	6
9	6	4	7
10	3	2	4
11	2	1	3
12	1	1	6
13	6	3	5
14	4	5	3
15	4	2	2
16	4	4	7
17	5	6	1
18	1	1	5
19	3	5	2
20	4	6	4
21	5	2	5
22	2	4	4
23	6	7	6
24	6	4	5
25	5	5	5
26	5	3	5
27	1	2	6
28	1	3	1
29	1	6	7
30	6	5	7
31	2	6	5
32	4	4	2
33	1	1	3
34	1	2	7
35	1	5	6
36	4	3	7

Question 8

Subject#	Neutral	Non-Cont.	Rules
1	5	5	3
2	7	7	7
3	4	4	1
4	6	6	7
5	2	2	4
6	1	4	4
7	5	6	6
8	1	5	5
9	4	3	4
10	3	2	4
11	1	3	5
12	1	1	4
13	5	2	5
14	5	5	5
15	2	1	2
16	2	4	6
17	3	5	1
18	1	1	5
19	1	4	2
20	2	5	1
21	3	1	3
22	1	2	6
23	7	7	5
24	1	1	1
25	2	3	3
26	5	4	4
27	1	1	5
28	1	2	1
29	1	5	7
30	4	3	4
31	1	6	3
32	3	3	2
33	1	1	2
34	3	2	6
35	1	2	4
36	4	1	5

Question 9

Subject#	Neutral	Non-Cont.	Rules
1	5	4	4
2	7	7	7
3	4	5	2
4	3	7	7
5	1	1	3
6	2	5	3
7	5	6	5
8	3	5	5
9	5	4	5
10	3	2	6
11	3	4	5
12	1	1	5
13	4	3	5
14	1	4	4
15	2	2	1
16	1	4	6
17	3	5	1
18	1	2	3
19	2	6	1
20	1	2	1
21	5	5	3
22	1	3	5
23	6	7	6
24	4	4	4
25	4	5	4
26	4	3	1
27	2	1	6
28	1	5	3
29	1	2	7
30	6	4	6
31	3	5	4
32	4	2	6
33	3	3	1
34	3	1	6
35	1	2	5
36	2	1	4

Question 10

Subject#	Neutral	Non-Cont.	Rules
1	4	5	5
2	7	7	7
3	6	4	1
4	6	6	7
5	3	1	3
6	1	5	3
7	6	6	5
8	4	6	5
9	6	4	7
10	4	2	4
11	3	4	5
12	1	1	3
13	4	3	4
14	1	4	4
15	2	1	3
16	4	5	5
17	6	5	1
18	1	1	3
19	4	6	1
20	2	6	2
21	6	4	5
22	2	4	6
23	6	7	5
24	5	2	5
25	5	5	5
26	3	2	2
27	1	1	3
28	1	7	6
29	5	5	6
30	6	6	6
31	2	5	4
32	4	3	3
33	3	1	3
34	3	2	6
35	1	2	5
36	3	2	6

Appendix F

Questionnaires used in the Gracie User Study

This appendix contains the questionnaires given to users after each of the three interactions with Gracie. Also, the final questionnaire that was given after the final interaction is provided.

Evaluation of a System for Informing Students about the Graduate School Option by Dialog:

Questionnaire

Subject ID_________ Run ID__________

For each of the following questions, please rate your agreement by circling a number.

Question	**Rating** Strongly Disagree						Strongly Agree
1. I felt I had a connection with the coordinator	1	2	3	4	5	6	7
2. I think the coordinator and I understood each other	1	2	3	4	5	6	7
3. The coordinator seemed willing to help	1	2	3	4	5	6	7
4. The coordinator seemed trustworthy	1	2	3	4	5	6	7
5. The coordinator seemed likable	1	2	3	4	5	6	7
6. My conversation with the coordinator seemed natural	1	2	3	4	5	6	7
7. I enjoyed the interaction with the coordinator	1	2	3	4	5	6	7
8. The coordinator was human-like	1	2	3	4	5	6	7
9. The coordinator was persuasive	1	2	3	4	5	6	7
10. I would recommend the coordinator to others	1	2	3	4	5	6	7

Comments:

Evaluation of a System for Informing Students about the Graduate School Option by Dialog:
Final Questionnaire

Subject ID_________

Please circle one answer for each question.

1. **<u>Which system did you prefer the most?</u>**

   System 1 System 2 System 3

2. **<u>Which system did you prefer second most?</u>**

   System 1 System 2 System 3

3. **<u>Which system did you prefer the least?</u>**

   System 1 System 2 System 3

Please answer the following free response questions.

1. **<u>Do you have any suggestions for improvements?</u>**

2. **<u>Do you have any other comments or suggestions pertaining to the experiment?</u>**

Curriculum Vitae

Jaime Acosta was born in El Paso, Texas on August 12, 1982. Born as the third of three children from Jose and Maria Dolores Acosta, Jaime pursued a Bachelors degree at the University of Texas at El Paso in the year 2000. During this time he spent two years working under Dr. Teller conducting research for IBM. Jaime co-published four research papers on the effectiveness and efficiency of analyzing memory hierarchy behavior of the IBM Power4 architecture by using samples from traces of running the TPC-C benchmark. In 2003, Jaime graduated with honors and began working as a government civilian at White Sands Missile Range. During this time, Jaime began his graduate studies at the University of Texas at El Paso.

While pursuing his Masters degree, Jaime was under the direction of Dr. Nigel Ward. He became a member of the Interactive Systems Group which is a research group focused on human-computer interaction. Jaime received his Master's degree in May 2007. This was not the end of his academic journey. Jaime continued under the direction of Dr. Nigel Ward and conducted important research focused on improving spoken dialog systems. His work has been published at many venues around the world.

In December 2009, while still working full time, Jaime is the first in his family to successfully obtain a Doctoral degree.