2012-01-01

Adding Within-Utterance Emotion Decay for More Human-Like Dialog

Michael Hans Durcholz
University of Texas at El Paso, mdurc531@gmail.com

ADDING WITHIN-UTTERANCE EMOTION DECAY

FOR MORE HUMAN-LIKE DIALOG

MICHAEL HANS DURCHOLZ

Department of Computer Science

APPROVED:

Nigel Ward, Ph.D., Chair

David Novick, Ph.D.

Stephen Crites, Ph.D.

Benjamin C. Flores, Ph.D.
Interim Dean of the Graduate School

Dedication

to my family

and friends

with love

and appreciation

ADDING WITHIN-UTTERANCE EMOTION DECAY FOR MORE HUMAN-LIKE DIALOG

by

MICHAEL HANS DURCHOLZ, B.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

MASTER OF SCIENCE

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2012

Acknowledgments

I would like to thank God and my family for the support they provided for me during this time of my life. I would like to especially thank Alexis Sein, and my mother and father for their assistance with pilot studies, their useful feedback, and the immense support and patience they gave me during this process.

I would like to deeply thank Nigel Ward for his guidance in exploring more about graduate-level research and for his assistance as an advisor. I thank David Novick and Stephen Crites for being on my committee and for their constructive feedback.

I thank Shreyas Karkhedkar, Alejandro Vega, and the rest of the Interactive Systems Group at The University of Texas at El Paso for their comments and suggestions.

I am grateful to Jaime Acosta for providing the initial code base for Gracie. His code base for Gracie and his documentation proved to be immensely useful when carrying out the coding modifications and the experiment in this study.

I would also like to thank Timo Baumann from the University of Hamburg for his suggestions and assistance in working with the acoustic parameters of MaryTTS.

This work is supported in part by NSF grant IIS-0415150.

Abstract

While spoken dialog systems have been used for commercial applications for several decades, most commercial spoken dialog systems provide only simple information exchange capabilities. Emotion synthesis in spoken dialog systems has become an active research area recently, and use of emotion-adaptive dialog systems has demonstrated improvements in user experience and rapport. This thesis seeks to improve how emotions are conveyed in dialog systems to enable robust emotional support that improves user experiences with dialog systems and models human speech characteristics more accurately than current dialog systems.

Prior work with Gracie (GRAduate Coordinator with Immediate-response Emotions), an emotion-adaptive dialog system, enabled system utterances that conveyed emotional qualities based on the perceived emotional state of the user. This feature improved the user experience with the dialog system through increased rapport. However, Gracie was able to convey only a constant emotion during a conversation turn, regardless of the length of the turn. This thesis extends Gracie to modify the emotional qualities of system utterances on a sub-turn level.

In the study carried out in this thesis, the emotional coloring was varied on a sub-turn level by linearly attenuating the emotional qualities so that they reached a neutral emotional state at the utterance end. Evaluation with 36 subjects showed that they significantly preferred conversing with the version of Gracie that supports sub-turn emotional coloring over the original Gracie. Subjects also tended to rate the sub-turn coloring Gracie system as more human-like than the original system.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

As spoken dialog systems become more common in everyday life, people increasingly depend on using these systems to accomplish tasks with organizations. However, many of these dialog systems are limited to information exchange [Acosta and Ward, 2011]. In human-to-human communication, information exchange is only one important aspect of communication; the emotion of the speaker (inferred from tone, pitch, and other vocal aspects) is another important aspect of human-to-human verbal communication [Kiesler et al., 1984]. Previous work with emotion in dialog systems indicated that users prefer to interact with dialog systems that incorporated some emotional qualities, in particular when a dialog system with a notion of emotions synthesizes an emotion based on an appropriate for the perceived emotion of the user [Acosta, 2009; Acosta and Ward, 2011]. However, these experiments suggested that performing emotional coloring only once over a long utterance resulted in a degradation in perceived conversational quality by users.

1.1 *Aims*

This thesis focuses on extending emotion-adaptive dialog systems to appear more human-like and preferable. This is accomplished by enabling emotion-adaptive dialog systems to perform emotional coloring on a sub-turn level through an emotional coloring attenuation function.

Communication accommodation theory (CAT) serves as a theoretical basis for this optimization. In CAT, humans use prosody (which describes the rhythm, tone, stress, and intonation of speech) and nonverbal behaviors to decrease social distance through convergence, or sometimes to increase social distance through divergence [Giles and Ogay, 2007]. As shown in the next chapter, the principles of CAT are present in multiple levels of communication, from

sub-turn to conversation-wide. Through greater application of this theory, utterances in dialog systems can become more human-like and because of this more human-like behavior, users may prefer to converse with the more human-like dialog systems over state-of-the-art dialog systems.

1.2 *Thesis Statement*

The main claims of this research are

1. A spoken dialog system that controls emotional coloring on a sub-turn level will appear more natural and human-like than a dialog system that does not.
2. A spoken dialog system that controls emotional coloring on a sub-turn level will be preferred over a dialog system that does not.

To test these claims, I extended Gracie (GRAduate Coordinator with Immediate-response Emotions) [Acosta, 2009], a spoken dialog system that enables emotional coloring, to provide support for sub-turn level emotional coloring and evaluated the improved version of Gracie with users.

In the rest of this thesis, I first present a survey of pertinent literature, including findings of a study involving the previous version Gracie. I then describe the experimental setup and the different versions of Gracie used. I report the results and implications of the experiment. Finally, I conclude with potential avenues of future work.

Chapter 2: Related Work

This thesis reviews research on emotion and dialog systems, communication accommodation theory, emotion representation and emotion synthesis in dialog systems, and prior work on the emotion-adaptive dialog system extended in this study (Gracie).

2.1 *Communication Accommodation Theory*

The psychological concept of communication accommodation theory (CAT) helps describe many actions individuals may take in conversations. CAT provides a framework for individuals in discourse to adjust social distance through convergence or divergence of their speech and non-verbal behaviors [Giles and Ogay, 2007]. In CAT, convergence leads to accommodation in a conversation, where an individual changes his or her communicative behavior until the communicative behaviors of the participants are similar [Giles and Ogay, 2007]. In contrast, divergence leads to accentuation of verbal and nonverbal communicative differences between the conversation participants, where an individual changes his or her communicative behavior to differ the communicative behavior of another conversation participant [Giles and Ogay, 2007]. Whether convergence or divergence is applied depends on whether the speaker wishes to accentuate similarities or differences of opinions, points of view, or other personal aspects between the conversation participants [Giles and Ogay, 2007]. In the context of emotion, CAT applications occur in interpersonal communication when an individual exercises convergence by conveying a complementary or "appropriate" emotion to another individual in order to establish or maintain rapport (e.g. a friend may convey empathy to another friend who is depressed) [Giles and Ogay, 2007].

Phenomena that CAT describes are not required to range over several conversation turns; CAT is also evident on a sub-turn level. One dominant example of CAT on a sub-turn level is during potential turn transition points within a speaker's turn. Specifically, pauses or "trailing off" (consisting of a gradually slower speaking rate with more rapid spacing between words) may occur within a turn to indicate that a harmonious turn change may occur [Maroni et al., 2008]. Similarly an individual may begin a conversation turn demonstrating convergence in his or her interpersonal behavior but may begin to diverge by the end of the turn to emphasize the concluding sentence of his or her turn. In addition, an individual may converge at the end of his or her turn to illustrate more of a common ground between conversation participants. For example, when multiple (more than two) individuals in a group conversation contend for holding the next conversation turn, the "winning" individual may express initial convergence in terms of tempo matching then diverge to his or her natural tempo as the turn progresses [Hayashi, 1990]. Therefore, sub-turn level CAT is important to model for systems that aim for some level of human-like conversation because humans apply CAT's principles over multiple levels of their conversational behaviors.

2.2 *Emotion Representation and Emotion Synthesis*

In emotion-related research, emotions are predominantly described in two different types of representations: discrete and dimensional. Discrete emotions are what people traditionally think of as emotions. They are emotions described using words of a language, such as anger, disgust, fear, embarrassment and others [Ekman, 1992]. While discrete emotions have the benefit of being recognizable by humans, automating direct recognition of a discrete emotion in voice has proven difficult. Most recent automated processes for emotion recognition have involved using a dimensional approach to emotions and, optionally, translating these dimensions to some discrete

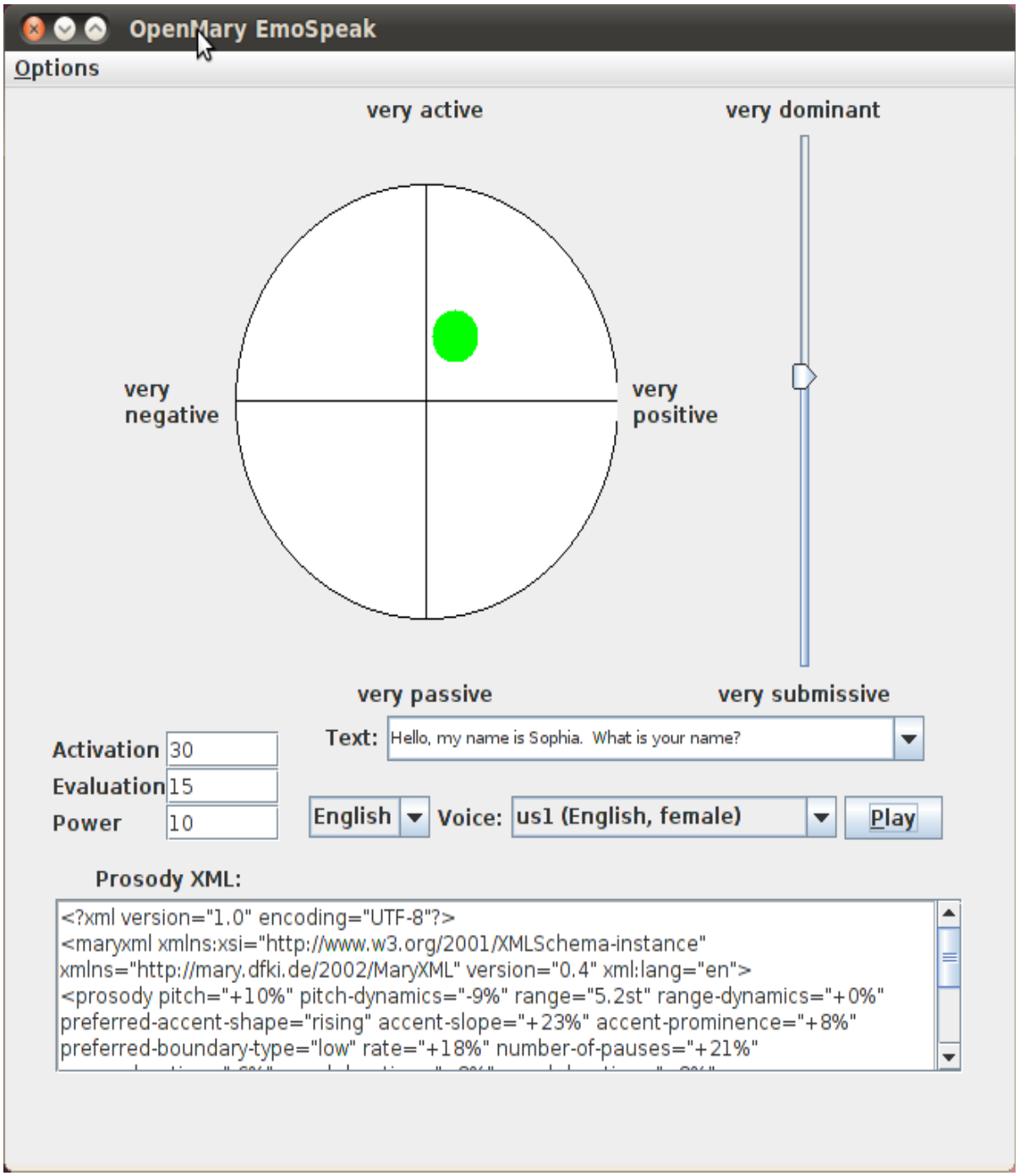

Figure 2.1: Screenshot of the EmoSpeak emotional coloring interface for MaryTTS

emotion when possible. This thesis uses the dimensional approach to emotion representation and synthesis and the term emotion is used broadly to also include attitudes and emotional states.

The dimensional approach to emotions has typically involved two to three dimensions. Researchers have described dimensional emotions using relative words [Schlosberg, 1954] or using numbers to quantify the dimensions [Bradley and Lang, 1994]. While different emotional dimensions have been used in previous emotion-related studies, most representations use dimensions representing some variants of valence (how positive/negative an individual is), activation (how active/passive an individual is), and power (how dominant/submissive an individual is). When these dimensions are represented using numbers, researchers typically use a number scale of -n to +n (such as -100 to +100). Using these dimensions, we can represent discrete emotions in the following way: depression possesses negative valence, anger possesses negative valence and (usually) positive activation, happiness possesses positive valence, etc.

Emotion synthesis within system-generated utterances has been feasible since Cahn's Affect Editor demonstrated that sentences can have different emotional colorings [Cahn, 1989]. With the Affect Editor humans were able to detect the emotions generated and classify them to discrete emotions [Cahn, 1989]. Emotion synthesis has also benefited from the dimensional representation of emotions as dialog systems can easily use this representation [Schröder, 2004b].

The dimensional representation of emotion formed a focal point of the dialog system interface used by MaryTTS. MaryTTS, an open-source speech synthesis engine and the speech synthesis engine used in this study, has an emotional coloring interface called EmoSpeak that provides emotional coloring based on three dimensions (activation, valence, and power) [Schröder and Trouvain, 2003]. A graphical user interface supplied with EmoSpeak (as seen in

Figure 2.1) enables users to graphically see how strongly or weakly the provided dimensional emotion values are in the scale of what EmoSpeak can provide [Schröder, 2004a]. EmoSpeak also provides an emotion dimension markup language that transforms the dimensional emotion values into the prosodic values of the utterance. EmoSpeak was evaluated for its effectiveness in conveying emotions to human users through a perceptual experiment. Testing of EmoSpeak demonstrated that conveying emotion through altering prosodic values of utterances showed correlations to how humans convey their emotions by altering elements of their speech [Schröder, 2004a]. This result enabled emotion-adaptive systems to specify rules translating dimensional emotion values to prosodic values to convey appropriate emotions to human users.

2.3 *Emotion-Adaptive Dialog Systems and Previous work with Gracie*

Work with emotion-adaptive dialog systems, where emotion is used to influence interaction, has been promising in terms of enhancing user experience. Students highly preferred conversing with an experimental tutoring-based dialog system that was designed to detect uncertainty through emotion and adapted system utterances based on the perceived uncertainty [Forbes-Riley and Litman, 2009]. Many potential domains exist for emotion-adaptive systems, one of which is academic career counseling. The remainder of this section will focus on previous work with Gracie (GRAduate Coordinator with Immediate-response Emotions), which this thesis extends.

Gracie is an emotion-adaptive dialog system that employed concepts of CAT by inferring the perceived emotional state of a student and the dialog system exercised (predominantly) convergence to establish common ground with the student and establish rapport with the student with the aims of giving the student useful information about graduate school. More specifically, Gracie analyzed the perceived dimensional representation of the person's emotion and generated a system response that was designed to exercise convergence by performing emotional coloring

on the system's responses to the user [Acosta, 2009; Acosta and Ward, 2011]. The dimensional representation of emotions used by Gracie were: activation (how active/passive the speaker is), valence (how positive/negative the speaker is), and power (how dominant/submissive the speaker is). The specified activation, valence, and power values were given to MaryTTS which modified the tone, pitch, and other prosodic elements of system's response [Acosta, 2009; Acosta and Ward, 2011]. The previous Gracie study compared the user-rated performance of two baseline emotionally-colored dialog systems (one system was non-contingent on the user's perceived emotion, and another dialog system incorporated a constant neutral emotion) against a dialog system that emotionally colored system utterances based on the user's perceived emotion. The test subjects noted an improved user experience and increase in rapport with the emotionally-contingent dialog system over the two baseline dialog systems.

However, Gracie was not designed to support all potential applications of CAT. In particular, Gracie did not support sub-turn level applications of CAT. In the context of Gracie, the dialog system was only able to perform emotional coloring that spanned the entire conversation turn; regardless of how long the conversation turn, the emotion values would remain constant [Acosta, 2009]. Psycholinguistic work on classification of emotion-related episodes showed that emotions can vary drastically within an utterance, particularly during a long conversation turn [Batliner et al., 2010]. In addition, as seen in the Communication Accommodation Theory section, humans can alter different prosodic elements of their utterances within a turn to serve some agenda (such as securing a future conversation turn during a long conversation).

So this version of Gracie was not able to implement and express a key principle of CAT that takes place in everyday human-to-human communication. It is important to incorporate the

capability to alter emotional values on a sub-turn level to incorporate more CAT principles in dialog systems and thus make dialog system conversations akin to “human-like” conversation. By allowing for emotional coloring on a sub-turn level, we can answer the following question: will a dialog system that emotionally colors words on a sub-turn level appear more “natural” and “human-like” (with respect to human-to-human speech) and be preferred to converse with by humans over a dialog system that does not color words on a sub-turn level?

2.4 *Summary*

Through incorporating elements of CAT, we can model more natural and human-like conversations by adding numerous points of introducing convergence and divergence in the conversation. It is important to note that different conversation topics or motives will affect whether convergence or divergence is dominant in the conversation. Nevertheless, CAT states that shifts in convergence and divergence application can occur within a conversation turn, and this should be modeled to achieve more human-like conversation.

While the discrete emotion representation is usually considered to be the most natural representation for humans, the dimensional representation of emotion shows promise for emotion-adaptive dialog systems and especially for conveying system-generated responses that are emotionally colored. Because of this representation and through inclusion of aspects of CAT, emotion-adaptive dialog systems such as Gracie have demonstrated an improved user experience. However, current emotion-adaptive dialog systems do not encompass all of CAT, namely the sub-turn applications of CAT. I hypothesize inclusion of these capabilities can further enhance user experience and more closely model human-like speech.

In this thesis, I used the findings from sub-turn level applications of CAT and previous work with emotion-adaptive dialog systems, particularly Gracie, to extend Gracie to support sub-

turn level manipulation of emotional qualities. This thesis shows that the sub-turn level support results in improved user experience and suggests that such support will assist dialog systems in incorporating more human-like speech patterns.

Chapter 3: Experimental Setup

The study outlined in this thesis followed a methodology similar to that of prior experiments involving Gracie, with some modifications to the dialog systems and to the experimental procedure.

3.1 *Conditions and Measures*

The research questions related to this study are to see if a human feels that a dialog system that emotionally colors on a sub-turn level is considered more "natural" and "human-like" and that they prefer conversing with such a system over a dialog system that does not emotionally color on a sub-turn level. A total of three different versions of Gracie are tested in this study: two control configurations of the dialog system and an experimental system that colors on a sub-turn level. The two control configurations were developed as part of a prior study involving Gracie (Non-Contingent and Constant Rule-Based, referred to as Rule-Based in prior work) [Acosta, 2009; Acosta and Ward, 2011]. Each of the three versions of Gracie in this study perform some form of emotional coloring on system utterances.

The following are more detailed descriptions of the three configurations:

- Non-Contingent: this Gracie version emotionally colors system utterances but does not do so based on the perceived emotional state of the current user. Instead, this configuration refers to the emotional state of the user who last interacted with the Constant Rule-Based version of the dialog system [Acosta, 2009; Acosta and Ward, 2011]. Thus, the emotional coloring performed during this interaction is a potentially-

valid emotional coloring but the coloring choices are not contingent upon the current user's emotions.

- Constant Rule-Based: this Gracie version emotionally colors system utterances based on the perceived emotional state of the current user. This Gracie version utilizes the same emotional rules as used in prior Gracie studies so that comparability analyses of this study with prior Gracie studies can be performed [Acosta, 2009; Acosta and Ward, 2011].

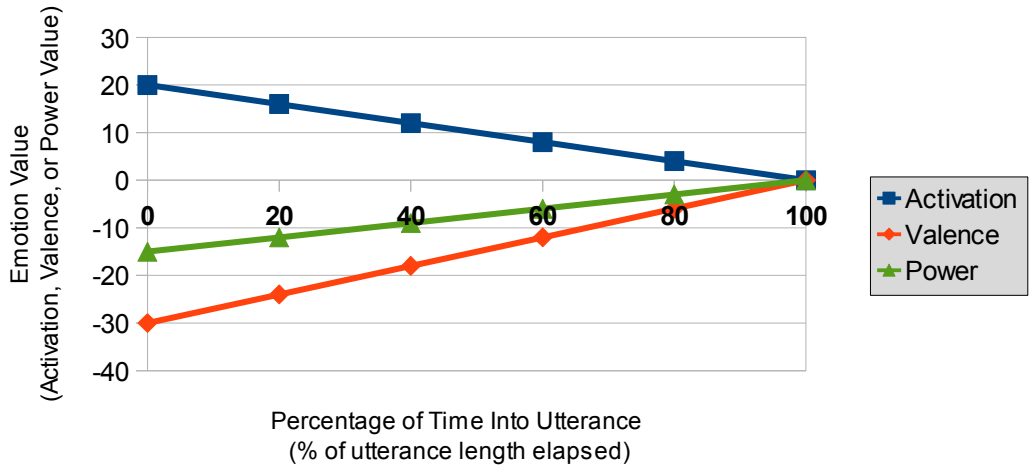

Figure 3.1: Example of how the linear-decayed rule-based system varies emotion on a sub-turn level

- Linear-Decayed Rule-Based: this Gracie version uses the same emotional coloring rules that the Constant Rule-Based version of Gracie does but applies them on a sub-turn basis. In addition, to implement more qualities of communication accommodation theory, this Gracie version incorporates the capability to change emotional qualities on a sub-turn level by “decaying” the emotional values to a “neutral” emotional state at the end of the

conversation turn (Activation/Valence/Power = 0 at the end of the utterance). An example of how the decaying occurs over the time into utterance is shown in Figure 3.1. The implemented attenuation function is a simple proof of concept that demonstrates usage of sub-turn level emotional value manipulations. The attenuation function is also described in section 3.3.3.

As in prior studies involving Gracie, each Gracie version was designed to be used with one of three content sequences. Because this is a within-subjects study, three different content sequences were used to minimize the effects of familiarity and user fatigue on the study from hearing the same content multiple times [Acosta, 2009]. In a pilot study, users had lengthy interactions with each Gracie version (>= 5 minutes), and users noted that user fatigue set in due to the long interaction length. The content sequences were shortened from prior experiments to further minimize user fatigue on the study, particularly since each content sequence included several system conversation turns that lasted up to 30 seconds each. The first sequence involved giving information about the statement of purpose, the second sequence's topic was about the Graduate Record Examination, and the third sequence expressed information about research and differences between undergraduate and graduate programs. The content sequences used are included in Appendix B.

3.2 *System Modifications*

To minimize the possibility of the test subjects having a negative user experience with Gracie due to something not directly related to the emotional coloring aspect of the systems (how frequent the coloring occurs and what coloring values are applied), several modifications were made to the dialog system. As in previous studies that involved Gracie [Acosta, 2009], Gracie

assessed the emotional state of the user by analyzing only the acoustic features in the user's voice, and the user's words are ignored. In addition, a static conversation flow from the system's standpoint was used to minimize the possibility of the dialog system misinterpreting what the user said. Because of the inclusion of both of these features, speech recognition was turned off within Gracie, since no decisions needed to be made based on the user's words.

Previous Gracie experiments introduced visual prompting to the user (the user would see the words that the dialog system was saying and would see a “Please Speak” prompt when the system was ready for user input) [Acosta, 2009]. In pilot studies of this Gracie experiment, synchronization issues sometimes occurred when ambient noise (such as deep breathing) was perceived as speech and would start Gracie's audio capturing process. To prevent this from occurring during the actual experiment, a “Push-to-Talk” style interface was implemented where the user would press and release the “Enter” key once prior to speaking with the system. This also enabled the user to think about their responses if time was needed.

3.3 *Linear-Decayed Rule-Based version of Gracie*

To enable manipulation of the Activation, Valence, and Power values within a conversation turn within the existing Gracie framework, extensive code changes were required in how Gracie communicated with MaryTTS and EmoSpeak within MaryTTS. The previous version of Gracie took the text that needed to be said in the next conversation turn from the content database, compiled a MaryXML based on the text to say combined with indicative prosodic values to apply to the entire conversation turn (the indicative prosodic values were determined by the emotion rules that modulated pitch, speaking rate, pausing, and other prosodic features of the dialog system's utterances), then sent this MaryXML to MaryTTS for EmoSpeak to parse through.

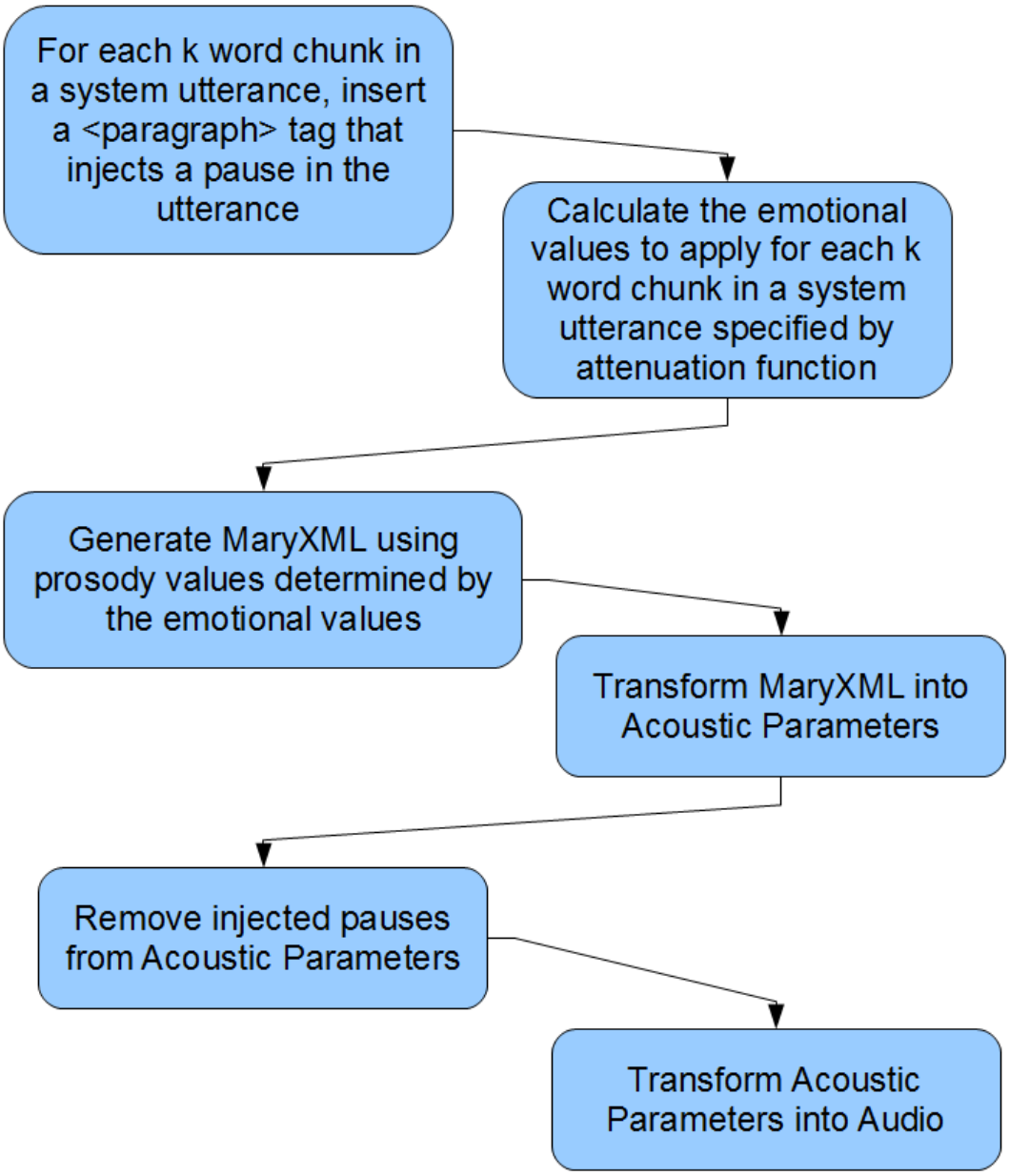

Figure 3.2: Algorithm followed by the linear-decayed rule-based system to support sub-turn level convergence based on emotional value manipulation. In this study, k = 2.

For this study, Gracie needed to be extended to enable application of the indicative prosodic value calculations on a sub-turn basis. One trivial approach to enabling this capability would be to generate multiple MaryXML "chunks" for each conversation turn, generating a MaryXML chunk for a contiguous set of words (e.g. set length = 2, 3, … words) that represents a sub-turn unit for which prosodic values would be assigned to. However, this initial, naïve approach was flawed for two reasons: the values specified in MaryXML files are merely indicative (the speech synthesis processor can decide to ignore the prosodic values specified) [Schröder and Trouvain, 2003], and undesirable pauses would be injected into the system utterances. The following subsections explain in detail the evolution of how sub-turn level prosodic value calculations were added into Gracie and how the flaws in the initial approach were addressed. Figure 3.2 specifies the algorithm used for the final approach implemented into Gracie to add sub-turn level emotional value manipulation.

3.3.1 Working Around the Indicative Nature of the Prosody Tag

The prosodic values specified in MaryXML files are indicative, which enables the speech synthesis processor to ignore prosodic values that would result in a perceived reduction in speech quality [Schröder and Trouvain, 2003]. However, for this study, the experimenter needed to ensure that the speech synthesis processor would not ignore the prosodic values passed to MaryTTS and would honor them exactly. Part of how the speech synthesis processor determines that a degradation of speech quality may occur is by observing how fast prosodic values change with respect to time. Knowing this, if one were to increase the span of time over which a prosodic value change occurs, the check would determine that less risk of degradation of speech quality occurred.

Therefore, a workaround needed to be implemented to add more time within the utterances so that the speech synthesis processor would not fear any degradation of speech quality. This led to using the <paragraph> or <p> tag within the MaryXML data surrounding each MaryXML chunk. This was a partial success in that MaryTTS explicitly obeyed the prosodic values specified within each chunk.

However, by generating multiple MaryXML chunks per conversation turn, delimited by <paragraph> tags, many unwanted pauses would be introduced to the system utterances (e.g. the system would speak with a cadence similar to: <word> <word> <word> <2 second pause> <word> <word> <word> <2 second pause> …). In pilot studies, this pausing pattern was seen to have a dramatic detrimental effect on user experience with the dialog system, with many users commenting that it "sounded like a robot" because of the unexpected pause after every chunk. Considering how a major aspect of this study was to try to approximate human-like speech as much as possible, the undesired pausing pattern was not preferred for the dialog system version used for the experiment. While the dialog system now explicitly obeyed the prosodic values (determined from emotion rules) specified in the MaryXML script, this study necessitated approximating human-like speech as much as possible, so these injected pauses needed to be removed. To remove the pauses, more in-depth knowledge of how MaryTTS worked was required.

3.3.2 Acoustic Parameters Manipulation

As MaryTTS has multiple steps in the procedure to turn a MaryXML script into system-generated audio, for the purposes of this study, it was necessary to find an intermediate step that had values that were not merely indicative but instead were explicitly obeyed by the speech synthesis processor. MaryTTS indeed has such an intermediate layer, called acoustic parameters,

which defines the durations and fundamental frequencies for each phoneme of the system's utterance, including any pauses that needs to be generated [Schröder and Trouvain, 2003]. From analysis of the acoustic parameters of MaryXML scripts that utilized the <paragraph> tag delimiter for chunks, the tag inserted a pause of a constant length. Using this knowledge, a script was developed that generated a MaryXML script that delimited chunks using the <paragraph> tags, sent this MaryXML to MaryTTS, and retrieved the output acoustic parameters from the MaryTTS request. Then the pauses inserted from the <paragraph> tags were removed from the acoustic parameters and the new version of the acoustic parameters was sent to MaryTTS to generate the resultant audio.

Hearing audio generated via the improved approach had a “less robotic” cadence and still followed the emotion and prosodic values specified. It was noted by individuals in the lab that some co-articulation was lacking at the chunk boundaries; however, the effect was not significant to untrained ears. This approach enabling sub-turn level emotional coloring was used for the three dialog system versions to ensure consistency of the stimuli (e.g. to minimize test subjects preferring either the Non-Contingent or Constant Rule-Based system simply because they noticed better co-articulation at the chunk boundaries). For these implementations, only the Linear-Decayed Rule-Based version utilized the sub-turn level capabilities of the new code design (the Non-Contingent and Constant Rule-Based versions held constant emotional values over each chunk).

3.3.3 Emotional Value Attenuation Function

The improved version of Gracie incorporated an emotional value attenuation function that modified the conveyed emotion of the system on a sub-turn level. In this study, a simple linear attenuation function was implemented that transited to the neutral emotion state at the end of a

conversation turn. This attenuation function was used because it is simple to implement and because it seems that humans generally tend to "dull" their emotions, or lose emotional distinctness, towards the end of a conversation turn in a "normal, non-aggressive" conversation. An example of the linear attenuation function in action is shown in Figure 3.1.

3.4 *Experimental Procedure*

The experimental procedure used in this study followed a similar methodology as what was presented in prior Gracie studies [Acosta, 2009]. First the test subjects were given an outline of what the experimental process entails and to expect the experiment to last around 15-20 minutes, with time provided after the experiment for any questions they may have. Before the experiment began, the test subjects were told that they would be filling out questionnaires about their experience with each dialog system version. The test subjects proceeded to read and sign a consent form and fill out a demographic sheet, which enables analysis based on gender, age, linguistic background dimensions.

After filling out the initial documents, the experimenter demonstrated a short interaction (2-3 conversation turns) with Gracie and outlined the limitations with the dialog system interactions. In the experimental setup, the test subjects held a headset to speak into the system; however, to enable the experimenter to intervene in case any audio skipping or detrimental system performance not correlated to the dimensions of the study occurred, the dialog system's audio came from the computer's speakers instead of the headset. Similar to previous Gracie studies [Acosta, 2009], the experimenter discussed the visual prompting (of the words and the "Please Speak" prompt) with the test subjects. In addition, the Non-Contingent configuration was prepared for each subject by copying the emotion response sequences from the previous user's interaction with the Constant Rule-Based system.

The experiment was counterbalanced to control for potential influences of familiarity playing a role in the experiment (e.g. the test subject preferring the last system simply because they felt more comfortable with the experimental setup on the last interaction). Each test subject received a unique permutation of system order and content sequence order. The test subjects were asked to evaluate their interactions with each version by filling out a 7-point Likert scale questionnaire on their experience with each system version. They were requested to fill out the questionnaire based on the dialog system's voice, not on the information or actual words conveyed. The questionnaire was based on the work from previous dialog system studies [Gratch et al., 2007; Acosta, 2009] to enable comparisons of the results from this study with the results from the prior Gracie study.

A question was added to the Likert scale questionnaire to assess Likert-scale based system preference: (#10) “I preferred talking to the coordinator over a human”. This additional question compared to a specified baseline (conversing with a human) to increase the reliability of the results from this question as well as normalize based on the test subjects' experiences with human-to-human communication (e.g. on a version of this question that does not compare to some baseline, an introvert may denote that they highly prefer conversing with the system because they may not like conversing with other people, which would not be represented in the results). In the context of the hypotheses of this study, questions #6, #8, and #10 of this questionnaire are most pertinent, as follows:

- Hypothesis #1: Test subjects will rate a spoken dialog system that controls emotional coloring on a sub-turn level as more “natural” and “human-like” on questionnaires than dialog systems that do not emotionally color on a sub-turn level:
    - Question #6: My conversation with the coordinator seemed natural.

- Question #8: The coordinator was human-like.

- Hypothesis #2: Test subjects will rate a spoken dialog system that controls emotional coloring on a sub-turn level as more "preferable" on questionnaires than dialog systems that do not emotionally color on a sub-turn level:
  - Question #10: I preferred talking to the coordinator over a human.

After completing each of the three interactions with the dialog system and the associated Likert scale questionnaires, the test subjects were asked to fill out a comparison-based questionnaire comparing their experiences with the three dialog system versions. A copy of the experimental procedure followed in the study is in Appendix A. In addition, the questionnaires that the participants filled out are in Appendix C.

3.5 *Subject Pool*

Thirty-six test subjects participated in this study, 30 of whom were male and 6 were female. All test subjects were students enrolled in an introductory computer science course (Introduction to Computer Science) at the time of the study. Most of the participants were computer science majors, while some were mathematics majors. As part of their enrollment in the Introduction to Computer Science course, the students needed to complete two research credits by participating in department events or research experiments such as this study; the students received one of their required research credits through their participation in this study. None of the students had interacted with any version of Gracie or with the graduate coordinator on which Gracie is based [Acosta, 2009; Acosta and Ward, 2011].

Chapter 4: Results

The results from this study suggests that the first hypothesis is true, that users think that a dialog system that controls emotional coloring on a sub-turn level is more “natural” and “human-like”, while the results of the study supports the second hypothesis, that users prefer conversing with a dialog system that controls emotional coloring on a sub-turn level. This chapter presents in detail these and other results as well as associated discussion.

4.1 *Rating Questions Analysis*

After each of the three dialog system versions (Non-Contingent, Constant Rule-Based, and Linear-Decayed Rule-Based), subjects filled out a the Likert scale questionnaire. The results from the rating questions were analyzed for any possible correlations (as shown in Table 4.1, discussed in section 4.3.1, and included in Appendix E) and the ratings were assessed for significance using a paired-sample one-tailed t-test. Table 4.2 shows the average ratings given for each question, the standard deviation for each question, and any significance determined from the t-test. The p values associated with t-tests are included in Appendix E. For the results in this thesis, statistical significance occurs when $p < 0.05$, and a statistical tendency occurs (e.g. a result is statistically suggestive) when $p < 0.10$.

Except for the Cognitive Rapport question, all of the results for the Linear-decayed Rule-Based version of the dialog system were higher than for both of the control systems (Non-Contingent and Constant Rule-Based).

The rating for the Linear-decayed Rule-Based system was statistically significant over the Non-Contingent system for the Human-like question. In addition, the ratings for the Linear-decayed Rule-Based system were statistically significant over the Constant Rule-Based system

Table 4.1: Correlation analysis on the rating question results.

Q #	**Scale**	**Q #**	**Scale**	**Q #**	**Scale**	**Q #**	**Scale**
1	Emotional Rapport	2	Cognitive Rapport	3	Helpful	4	Trustworthy
5	Likeable	6	Natural	7	Enjoyable	8	Human-like
9	Persuasive	10	Preference	11	Recommendable		

Q #	1	2	3	4	5	6	7	8	9	10	11
1	1	0.82	0.72	0.7	0.71	0.73	0.66	0.59	0.67	0.44	0.77
2	0.82	1	0.69	0.65	0.68	0.74	0.59	0.56	0.58	0.29	0.66
3	0.72	0.69	1	0.85	0.79	0.64	0.69	0.51	0.7	0.27	0.66
4	0.7	0.65	0.85	1	0.86	0.68	0.77	0.57	0.79	0.25	0.67
5	0.71	0.68	0.79	0.86	1	0.72	0.78	0.7	0.79	0.33	0.65
6	0.73	0.74	0.64	0.68	0.72	1	0.68	0.73	0.68	0.34	0.66
7	0.66	0.59	0.69	0.77	0.78	0.68	1	0.66	0.76	0.44	0.68
8	0.59	0.56	0.51	0.57	0.7	0.73	0.66	1	0.72	0.46	0.65
9	0.67	0.58	0.7	0.79	0.79	0.68	0.76	0.72	1	0.46	0.82
10	0.44	0.29	0.27	0.25	0.33	0.34	0.44	0.46	0.46	1	0.49
11	0.77	0.66	0.66	0.67	0.65	0.66	0.68	0.65	0.82	0.49	1

This supplementary table describes correlation values for questions pertaining to hypotheses.

Natural ↔ Human-like Correlation	0.73
Likeable ↔ Enjoyable Correlation	0.78
Likeable ↔ Preference Correlation	0.33
Enjoyable ↔ Preference Correlation	0.44
Natural ↔ Preference Correlation	0.34
Human-like ↔ Preference Correlation	0.46

for the Likeable, Enjoyable, and Preference questions. Finally, the rating for the Linear-decayed Rule-Based system was statistically suggestive over the Constant Rule-Based system for the Human-like question.

4.1.1 Hypothesis #1 Analysis

In the context of the first hypothesis, the Natural and Human-like rating questions were analyzed. For the first hypothesis (test subjects will rate the Linear-Decayed Rule-Based system as more "natural" and more "human-like" than the Constant Rule-Based system), while the Linear-Decayed Rule-Based system received higher scores on both questions #6 and #8, only the "human-like" score (question #8) is statistically suggestive ($p = 0.06$). With respect to hypothesis #1, question #8 is likely more relevant to the initial aims of the study because part of the study's motivation involves developing more "human-like" interactions with dialog systems. Therefore, hypothesis #1 is partially supported and further work should investigate this.

4.1.2 Hypothesis #2 Analysis

In the context of the second hypothesis, the Preference rating question was analyzed. With respect to the second hypothesis (test subjects will rate the Linear-Decayed Rule-Based system more "preferable" than the Constant Rule-Based system), the Linear-Decayed Rule-Based system received a higher score that was statistically significant ($p < 0.02$). In addition, the Linear-Decayed Rule-Based score is significantly higher than the Non-Contingent score ($p < 0.01$). Therefore, hypothesis #2 is supported based on the rating question results.

4.1.3 Rating Question Comparison with the Previous Gracie Study

One of the aims of this study was to replicate the results from the previous Gracie study; however, the results from this study do not confirm these results. Unlike in prior work with

Table 4.2: Subjects' ratings of the three versions of Gracie

+++ – higher than Non-Contingent, p<0.02, *** – higher than Constant Rule-Based, p<0.02

++ – higher than Non-Contingent, p<0.05, ** – higher than Constant Rule-Based, p<0.05

+ – higher than Non-Contingent, p<0.10, * – higher than Constant Rule-Based, p<0.10

		Means (Standard Deviations)		
Q #	**Scale**	**Non-Contingent**	**Constant Rule-Based**	**Linear-decayed Rule-Based**
1	Emotional Rapport	4.5 (1.61)	4.56 (1.52)	4.67 (1.43)
2	Cognitive Rapport	4.67 (1.62)	4.81 (1.86)	4.53 (1.36)
3	Helpful	5.39 (1.63)	5.47 (1.70)	5.53 (1.80)
4	Trustworthy	5.31 (1.43)	5.25 (1.61)	5.33 (1.66)
5	Likeable	* 5.28 (1.56)	5.00 (1.67)	*** 5.42 (1.60)
6	**Natural**	4.22 (1.64)	4.33 (1.80)	4.39 (1.68)
7	Enjoyable	* 5.25 (1.64)	4.97 (1.84)	** 5.33 (1.60)
8	**Human-like**	3.75 (1.52)	3.81 (1.75)	++, * 4.11 (1.55)
9	Persuasive	4.42 (1.66)	4.44 (1.95)	4.66 (1.59)
10	**Preference**	2.92 (1.52)	3.09 (1.78)	+++, *** 3.47 (1.65)
11	Recommendable	4.39 (1.61)	4.33 (1.76)	4.56 (1.52)

Gracie [Acosta, 2009; Acosta and Ward, 2011], the Constant Rule-Based system ratings for the Trustworthiness, Likeable, Enjoyable, and Recommendable questions were lower than the ratings for the Non-Contingent system, with the Likeable and Enjoyable ratings for the Non-Contingent system being statistically suggestive over the Constant Rule-Based system. In addition, the Constant Rule-Based ratings for the Emotional Rapport and Cognitive Rapport questions were not significantly nor suggestively higher than the Non-Contingent ratings in this

Table 4.3: Subjects' system preferences and naturalness judgments for the three versions of Gracie

System	Most Preferred	2nd Most Preferred	Least Preferred	Most Natural	2nd Most Natural	Least Natural
Non-Contingent	11	13	12	12	8	16
Constant Rule-Based	12	13	11	11	14	11
Linear-Decayed Rule-Based	13	10	13	13	14	9

study, while the Constant Rule-Based ratings for these questions in the previous Gracie study were significantly higher.

4.2 *Comparison-Based Questionnaire Analysis*

The results from the comparison-based questionnaire are described in Table 4.3. In the comparison-based questionnaires, the test subjects slightly preferred conversing with the Linear-Decayed Rule-Based system and felt that the system was slightly more natural. However, it is critical to note that none of the results from the comparison-based questionnaire were statistically significant or suggestive via chi-squared tests. This could be partially due to aspects of the testing procedure: the test subjects had three interactions with dialog system versions and had to recollect their experiences with each version while filling out the final questionnaire. While filling out the comparison-based questionnaire, some of the test subjects verbally stated that they had difficulty remembering some details with their earlier interactions while some other test

subjects stated that they may have gotten some interaction details confused between dialog system versions while filling out this final questionnaire. Therefore, none of the results from this questionnaire neither support nor refute any of the hypotheses, and the comparison-based results are considered inconclusive. In future studies, testing modifications should be made to increase the relevance of the comparison-based questionnaire. The chi-squared test results are included in Appendix E.

4.3 *Implications and Discussion of the Results*

The findings of this study suggest that there are some user experience benefits through the inclusion of sub-turn level emotional coloring. In addition, the study suggests that part of the improvement in user experience/preference with the improved Gracie system was due to Gracie's communicative behavior appearing more human-like. However, it is important to note that the significance of the benefit through adding sub-turn level emotional coloring is not as great as the benefit of adding emotionally coloring utterances. This is somewhat expected because the sub-turn level emotional coloring is an extension of generic emotional coloring capabilities and builds on these capabilities. Nevertheless, this extension can provide new capabilities to dialog system developers and may enable them to enhance user experience relatively easily (as discussed in section 5.2).

4.3.1 Rating Question Correlations

I assessed for correlations between the rating questions to determine if any redundancy was present in the rating questions. Table 4.1 includes both the correlations between rating question and also correlation values between Likeable, Natural, Enjoyable, Human-like, and Preference rating question results. For this thesis, a very strong correlation is present when the correlation

value is >= 0.7, a strong correlation is present when the correlation value is >= 0.5, and a weak correlation is present when the correlation value is < 0.5.

From the findings in Table 4.1 and based on the rating question data, a very strong correlation occurs between the Natural and Human-like questions. Another important finding of the correlation table is that no strong correlation exists between the Preference rating question and any other rating question. This finding shows that the Preference results are not explained by any other single factor, although the Preference results have "high weak correlation" values with all the other rating questions. The correlation values can be used in future Gracie studies to potentially reduce the number of rating questions on the questionnaires and make these questionnaires more robust.

4.3.2 Discussion of Rating Question Result Differences with the Previous Gracie Study

The discrepancies in statistical significance between this study and the previous Gracie study (as discussed in section 4.1.3) could mean that the Constant Rule-Based system is not significantly better in these aspects contrary to the previous result. Further work should attempt to re-examine the validity of the significance seen in the previous Gracie study.

It is possible that the discrepancies between this study and the previous Gracie study could be due to slight experimental differences. Comparing the two studies, shortened versions of the scripts from the previous study were used by the dialog systems in this study and the overall demographics of the test subjects in this study differed from the previous study. The number of system conversation turns used in this study were shortened to reduce effects of user fatigue because each system conversation turn was longer. This may have reduced the variation of the user experience partially because the users had less time to become accustomed to conversing with each dialog system version.

The fact that this study had a more homogeneous test subject population could have directly affected the results observed from the Non-Contingent system. The prior Gracie study had a test subject population of 23 males and 13 females and there was a greater variation in the age ranges of the test subjects [Acosta, 2009]. In this study, there were 30 males and 6 females and most of the test subjects were around age 18-24. Because the Non-Contingent system emotionally colors based on the emotional state of the previous user, the user experience is highly influenced by similarities and differences in behavioral patterns between the two users. Because this study had test subjects with more similar demographic information than the prior Gracie study, it is more likely that the test subjects in this study had similar behavioral patterns and therefore the Non-Contingent system likely produced emotional coloring that was deemed more appropriate to the current user.

It is important to note that the rating question results from this study are similar to previous Gracie studies with respect to Hypothesis #1 (questions #6 and 8) in that the Constant Rule-Based system received higher scores than the Non-Contingent system. However, the differences between the mean rating scores for the two questions from the previous Gracie study are much higher than the differences in means in this study (Question #6: 0.36 in prior study, 0.1 in this study; Question #8: 0.64 in prior study, 0.06 in this study) [Acosta, 2009]. Part of this could be due to the reasons outlined earlier in this section or also due to the inclusion of the Linear-Decayed Rule-Based system and the effects of this system version on user experiences.

4.3.3 Is "Human Natural" Not Equal to "Machine Natural"?

When looking at the rating questions results pertaining to Hypothesis #1, a noticeable disparity occurs between questions #6 and #8. Question #6 (the Natural score) is not statistically significant or suggestive ($p = 0.42$) while Question #8 (the Human-like score) is statistically

suggestive ($p = 0.06$). In addition, the difference in the mean ratings given for the Constant Rule-Based and Linear-Decayed Rule-Based systems is five times larger for Question #8 than for Question #6 (Difference in mean scores for Natural question = 0.06; Difference in mean scores for Human-like question = 0.30). Although there is a strong correlation in ratings as discussed in section 4.3.1, the difference in significance leads to an interesting question: with respect to dialog systems, do humans feel that "human-like" conversation with a dialog system is not "natural"? In other words, with respect to communication, is "human natural" not equal to "machine natural"?

While the question of "human natural" vs "machine natural" could have affected the results given by question #6 (Naturalness), it is important to note that the test subjects may have felt that Question #6 was ambiguous. Because the wording of the question "My conversation with the coordinator seemed natural" was not compared to an explicit baseline, the test subjects may have assessed the naturalness of the conversation with different baselines (such as prior conversations with humans, dialog systems, or other previous interactions). This ambiguity could have affected the accuracy of the Natural rating question results and researchers conducting future Gracie studies should consider rewording this rating question.

4.3.4 Relation of Results to Communication Accommodation Theory

The findings from this study relate to previous communication accommodation theory (CAT) studies through the Linear-Decayed Rule-Based system's design to more closely model sub-turn level human communicative behavior. From the results of rating questions #8 and #10, the Linear-Decayed Rule-Based system received higher scores than the control systems that did not model sub-turn level communicative behavior. As stated earlier for the Linear-Decayed Rule-

Based system, the human-like assessment rating question result was statistically suggestive and the preference rating question result was also statistically significant.

With respect to CAT, participants try to establish rapport in conversations by exercising convergence (e.g. conveying an “appropriate” emotion) with the implicit goal of having a more preferable and pleasing conversation with other conversation participants [Giles and Ogay, 2007]. In addition, most human-to-human communication incorporates some sub-turn level occurrences of convergence and divergence [Hayashi, 1990; Maroni et al., 2008]. Given that the goal of this study was to model more human-like and preferable communicative behavior, the improved version of Gracie applied convergence on a sub-turn level as opposed to just one convergence application for the entire utterance. The results of this study suggest that the sub-turn level application of convergence within Gracie generated more human-like and preferable communicative behavior.

4.3.5 Experimental Difficulties Experienced

While the results of the rating questions in this study are promising, difficulties and unexpected results occurred at times due to the nature of the experimental design. While the users were instructed to rate and comment on the systems based only on the quality of the voice produced and not the content that the system produced, many user comments revolved around the content conveyed during the interactions. The experiment used different system/content permutations to counterbalance and minimize these effects. However, some test subjects may have followed the instructions of only analyzing the voice quality while other test subjects may not have, which would decrease the effectiveness of the counterbalancing and increase the effect the content had on the results.

Participant judgments may have also been affected by a side effect of adaptation: increases in speaking rate in dominant, positive, or engaged contexts potentially leads to decreases in intelligibility and perceptions of voice quality. This is evident in an unfortunate aspect of the implementation of the Constant Rule-Based version: whenever a user speaks at a slower speaking rate, the Constant Rule-Based system moderates its speaking rate comparatively well and no decreases in voice quality and intelligibility are perceptible. When a user speaks at a normal or fast speaking rate, the Constant Rule-Based system's speaking rate increases, particularly in long utterances, and a decrease in voice quality and intelligibility is perceptible.

Chapter 5: Future Work

In this thesis, I showed that providing a finer granularity of where emotional coloring can be applied within an emotion-adaptive dialog system has a positive impact on user experience and suggested that dialog systems which include this can be perceived to be more human-like. Like prior work with Gracie, this work is just one step in improving the quality of dialog systems and in modeling human speech patterns and behaviors. Improvements to several aspects of the dialog system and associated experiments can be made to strengthen future emotion-adaptive dialog systems and to answer new psycholinguistic questions based on these systems.

5.1 *Improvements to the Current Experiment*

While the results from this study were positive, numerous improvements can be made to future experiments to increase the significance of the results and help generalize them further. One area of improvement involves increasing the significance and reliability of the comparison-based questionnaire results. Considering how many test subjects verbally stated that they had difficulty recalling qualities of each of the three interactions, a future study could involve playing back their conversations with Gracie to the user prior to filling out this questionnaire. This inclusion could help refresh the test subject memories of their experiences with each Gracie version, along with any perceived difficulties they had when conversing with the system. While this may have a detrimental effect on the test subjects' ratings on the questionnaire due to user fatigue with the study, user fatigue can be greatly minimized if the interactions are kept short in length. Adding questions to the comparison-based questionnaire regarding which system the test subjects felt was most/least human-like would help improve the relevance of this questionnaire on the results of future studies.

Another improvement to this study could involve test subject selection: the results were possibly affected by the relatively homogeneous test subject population. Compared with prior Gracie studies [Acosta, 2009], most of the test subjects were computer science students (all of the test subjects were either computer science students or mathematics students enrolled in a computer science course). It is possible that the comparison-based questionnaire results were hampered because the students were of a technical background, as opposed to liberal arts students (e.g. it is possible that computer science students may have a harder time distinguishing fine levels of emotion differences than psychology students).

5.2 *Avenues for Future Research*

Other research questions arise based on the results from this study. One potentially fruitful avenue of research would be to explore whether or not humans feel that a "human-like" conversation with a machine is not a "natural" conversation. This can be determined as a function of user expectations, past experiences and conversations, and/or compared with some baseline (e.g. the emotionally neutral system in prior Gracie studies [Acosta, 2009; Acosta and Ward, 2011]). Likert scale questions can also be added to the questionnaire asking how "machine natural" and "human natural" did the conversation feel.

Further work with emotional coloring attenuation functions can be conducted to more closely approximate human-like speech and to observe the effects on user experience of new attenuation functions. In the persuasion corpus (a human-to-human corpus that Gracie is based on), the dominance/power dimension values often decreased below the emotionally neutral state (below 0) by the end of the utterances, likely as a cue that a turn-taking opportunity was imminent [Ward and Escalante-Ruiz, 2009]. Further work can also explore applying different emotional coloring attenuation functions depending upon either local dialog state or some dialog

action, and applying hypotheses formed from meta-dialog analysis into future emotion-adaptive dialog systems [Ward and Vega, 2012]. Future work can also model emotion expression with respect to time and relate this to underlying mental processes that deal with emotion [Becker et al., 2004; Bosse et al., 2010]. Additionally, we can try versions of Gracie with different attenuation functions to see which one enhances user experience the most and which one test subjects think is most natural or human-like.

Another avenue of research could limit test subject sub-populations to individuals of particular demographics (male/female, ethnicity, age, college major, native/non-native speaker, etc.) and explore the preference/naturalness/human-like judgments these individuals make on the dialog system versions. Because applications of convergence and divergence in a conversation partially depend on characteristics (such as demographics) of the conversation participants [Giles and Ogay, 2007], a future study focused on analyzing the system ratings given by sub-populations could prove useful in optimizing future emotion-adaptive dialog systems. Tangentially, a Spanish-speaking version of Gracie could be developed (where the corpus data is Spanish and a Spanish voice is used with MaryTTS) and researchers can measure if the ratings and comparison-based judgments differ across linguistic lines or not.

5.3 *Summary*

Building upon the previous work with Gracie, this thesis showed that modulating the conveyed emotion on a sub-turn level has the ability to enhance user experience and suggested that more human-like conversation can be achieved through inclusion of this feature. The ability to modulate emotional qualities of system utterances enabled deeper rapport with dialog systems and the ability to modulate on a sub-turn level further extends the applicability and usefulness of emotion-adaptive dialog systems. The extension provided in this thesis also can provide more

flexibility for commercial dialog system production and can help bring emotion-adaptive dialog systems closer to widespread use in industry.

References

[1] Acosta, J. “Using Emotion to Gain Rapport in a Spoken Dialog System,” Ph.D. Thesis, University of Texas at El Paso, 2009.

[2] Acosta, J., Ward, N. “Achieving Rapport with Turn-by-Turn, User Responsive Emotional Coloring,” *Speech Communication*, 53(9-10):1137-1148, Elsevier, 2011.

[3] Batliner, A., Seppi, D., Steidl, S., Schuller, B. “Segmenting into Adequate Units for Automatic Recognition of Emotion-Related Episodes: A Speech-Based Approach,” in *Advances in Human-Computer Interaction*, Hindawi, 2010.

[4] Becker, C., Kopp, S., Wachsmuth, I. “Simulating the Emotion Dynamics of a Multimodal Conversational Agent,” in Andre, E. et al. [Eds], *Affective Dialogue Systems*, 154-165, Springer Verlag, 2004.

[5] Bosse, T., Gratch, J., Hoorn, J., Portier, M., Siddiqui, G. “Comparing Three Computational Models of Affect,” in Demazeau, Y., Dignum, F, Corchado, J, Perez, J. [Eds], *Advances in Practical Applications of Agents and Multi-Agent Systems*, 175-184, Springer Verlag, 2010.

[6] Bradley, M., Lang, P. “Measuring emotion: The self-assessment manikin and the semantic differential,” in *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49-59, Elsevier, 1994.

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

[8] Ekman, P. “An Argument for Basic Emotions,” in *Cognition and Emotion*, 6(3-4):169-200, Psychology Press Ltd, 1992.

[9] Forbes-Riley, K., Litman, D. “Adapting to student uncertainty improves tutoring dialogues,” in Proceedings of the International Conference on Artificial Intelligence in Education, IOS Press Amsterdam, 2009.

[10] Giles H. and Ogay, T. “Communication Accommodation Theory,” in Whaley, B. and Samter, W. [Eds], *Explaining Communication: Contemporary Theories and Exemplars*, 293-310, Lawrence Erlbaum Associates, 2007.

[11] Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R. “Creating rapport with virtual agents,” in Intelligent Virtual Agents: 7th International Conference, 125-138, Springer Verlag, 2007.

[12] Hayashi, R. “Rhythmicity, sequence and syncrony of English and Japanese face-to-face conversation,” in *Language Sciences*, 12(2-3):155-195, Elsevier, 1990.

[13] Kiesler, S., Siegel, J., McGuire, T. “Social Psychological Aspects of Computer-Mediated Communication,” in *American Psychologist*, 39(10):1123-1134, American Psychological Association, 1984.

[14] Maroni, B., Gnisci, A., Pontecorvo, C. “Turn-taking in classroom interactions: Overlapping, interruption, and pauses in primary school,” in *European Journal of Psychology of Education*, 23(1):59-76, Springer Verlag, 2008.

[15] Schlosberg, H. “Three dimensions of emotion,” in *Psychological Review*, 61(2):81-88, American Psychological Association, 1954.

[16] Schröder, M. “Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions,” in *Affective Dialogue Systems: Tutorial and Research Workshop*, 209-220, Springer Verlag, 2004a.

[17] Schröder, M. “Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis,” Ph.D. Thesis, Institut für Phonetik, Universität des Saarlandes, 2004b.

[18] Schröder, M., Trouvain, J. “The German text-to-speech synthesis system Mary: A tool for research, development and teaching,” in *International Journal of Speech Technology*, 6(4):365-377, Springer Verlag, 2003.

[19] Ward, N.G., Escalante-Ruiz, R. “Using subtle prosodic variation to acknowledge the user's current state,” Interspeech, 2009.

[20] Ward, N.G., Vega, A. “A Bottom-Up Exploration of the Dimensions of Dialog State in Spoken Interaction,” submitted to Sigdial 2012.

Appendix A: Experimenter Steps for the User Study

This appendix contains the steps the experimenter used in the testing procedure for the user study. The steps were adapted from the VoiceXML study conducted as a precursor to the prior Gracie study [Acosta, 2009].

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. Ensure TTS server and MySql database service are running

   b. 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 and fill out a questionnaire for each configuration

   d. fill out a final 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 questionnaires

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 the users will be interacting with the system

    a. Three different system configurations

    b. The system will respond to user vocal input

12. Give user a questionnaire after each configuration they converse with.

13. Give user the final questionnaire after all configurations are conversed with.

Closing

14. After the interactions with the system are complete, look the questionnaires over, and ask them follow-up questions about anything which is unclear or interesting. Write down key points of their responses in the margins.

15. Briefly explain the aims of the research; and answer any questions.

16. Note down any interesting questions or comments that came up.

17. Briefly explain how we'll use their data

a. we want to see which system configuration they thought was more persuasive/natural/pleasant to converse with.

b. we want to determine any shortcomings of the system and how we can improve

18. Promise to tell their TA that they participated.

19. Thank them warmly.

Appendix B: Dialog System Scripts Used for the User Study

This appendix contains the content that was used by the dialog systems in the user study. A total of three different scripts were used. None of the words said in each script were dependent on the dialog system version used in the interaction.

Script #1 Content:

- Hi, I'm Sophia, what is your name?
- Nice to meet you. I am here to help give you information about graduate school, what graduate school entails, and how one would go about applying for graduate school. Have you thought about graduate school yet?
- There are a couple of requirements for graduate school. There is a statement of purpose, an entrance test, and your grades to think about. It may seem like a lot but it's not.
- 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 going to consider your admission for the graduate program. These individuals should say that they want you in their program from your statement of purpose and other application materials. You're not gonna 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.
- If you have any further questions regarding graduate school, the application process, or how to write a successful statement of purpose, feel free to ask any faculty members or graduate students to gain their insight on their experiences.

Script #2 Content:

- Hi, I'm Sophia, what is your name?
- Nice to meet you. I am here to help give you information about graduate school, what graduate school entails, and how one would go about applying for graduate school. Have you thought about graduate school yet?
- There are a couple of requirements for graduate school. There is a statement of purpose, an entrance test, and your grades to think about. It may seem like a lot but it's not.
- 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 and 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 the admissions people decide if you're, you know, if you look like you're going to be a good match for graduate school.
- If you have any further questions regarding graduate school, the application process, or how to write a successful statement of purpose, feel free to ask any faculty members or graduate students to gain their insight on their experiences.

Script #3 Content:

- Hi, I'm Sophia, what is your name?
- Nice to meet you. I am here to help give you information about graduate school, what graduate school entails, and how one would go about applying for graduate school. Have you thought about graduate school yet?
- While graduate school has some college courses in common with undergraduate programs, for the most part graduate courses are more research-intensive and cover topics more in depth than undergraduate courses do. In addition, graduate courses usually have a term project where a student will have to complete a research paper after performing some combination of a literature survey of previous conference papers and a new code solution that the student is proposing to solve a problem.
- 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.
- If you have any further questions regarding graduate school, the application process, or how to write a successful statement of purpose, feel free to ask any faculty members or graduate students to gain their insight on their experiences.

Appendix C: Questionnaires Used

This appendix contains the questionnaires given to the users as part of this study. The first questionnaire was given after each interaction with a version of the dialog system. The second questionnaire is the comparison-based questionnaire given after all three interactions were completed.

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 preferred talking to the coordinator over a human	1	2	3	4	5	6	7
11. 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. **Which system did you prefer the most?**

System 1 System 2 System 3

2. **Which system did you prefer second most?**

System 1 System 2 System 3

3. **Which system did you prefer the least?**

System 1 System 2 System 3

4. **Which system did you think was most natural to converse with?**

System 1 System 2 System 3

5. **Which system did you think was second most natural to converse with?**

System 1 System 2 System 3

6. **Which system did you think was least natural to converse with?**

System 1 System 2 System 3

Please answer the following free response questions.

1. **Do you have any suggestions for improvements?**

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

Appendix D: Data from the User Study

This appendix contains the data collected during the user study. First the demographic information of the test subjects are described, associated with their subject ID number. This is followed by the order of the system version and conversation script they received, the scores they gave for the rating questions, and their comparative preference and naturalness assessment for the three system versions.

Subject #	Age	Gender	Occupation	Languages	Starting Age of Speaking Fluency	Country Language was Learned In
1	18-19	F	Student	English, Spanish	2, 2	U.S., U.S.
2	20-24	F	Student	English, Spanish, Japanese	2, 5, 5	U.S., U.S., U.S.
3	25-29	F	Student	English	1	U.S.
4	18-19	M	Student	Spanish, English	2, 5	Mexico, Mexico
5	18-19	M	Student	English, Spanish	3, 3	U.S., U.S.
6	25-29	M	Systems Administrator/Student	English, Spanish	4, 4	U.S., U.S.
7	20-24	M	Student	Spanish, English, French	0, 12, 16	Mexico, Mexico, Canada
8	30-34	M	Student	English	2	U.S.
9	18-19	M	Student	Spanish, English	2, 10	U.S., U.S.
10	18-19	F	Student	English, Spanish	10, 4	U.S., Mexico
11	18-19	M	Student	English	2	U.S.
12	18-19	F	Student	English	2	U.S.
13	20-24	M	Student	Spanish, English	0, 5	Mexico, Mexico
14	18-19	F	Student	Spanish, English	1, 5	Mexico, Mexico
15	20-24	M	Student	Spanish, English, French	6, 17, 18	Mexico, Canada, France
16	18-19	M	Student	English, Spanish	2, 2	U.S., U.S.
17	20-24	M	Student	Spanish, English	2, 12	Mexico, U.S.
18	18-19	M	Student	Spanish, English	2, 2	U.S., U.S.
19	18-19	M	Student	English, Arabic, Spanish	2, 2, 2	U.S., U.S., U.S.
20	18-19	M	Student	English	3	U.S.
21	40-49	M	Student	English	2	U.S.
22	40-49	M	Systems Engineer	Spanish, English	3, 3	U.S., U.S.
23	20-24	M	Warehouse/Student	English, Spanish	0, 5	U.S., U.S.
24	30-34	M	Web Developer/Student	English, Spanish	2, 2	U.S., U.S.
25	25-29	M	Student	English	2	U.S.
26	20-24	M	Student	English	2	U.S.
27	18-19	M	Assets Protection/Student	English	2	U.S.
28	20-24	M	Student/Technical Representative	Spanish, English	1, 5	U.S., U.S.
29	18-19	M	Retail associate/Student	English, Spanish	2, 2	U.S., U.S.
30	18-19	M	Student	English	4	U.S.
31	18-19	M	Student	English	2	U.S.
32	18-19	M	Cashier/Student	English, Spanish	3, 3	U.S., U.S.
33	18-19	M	Student	Spanish, English	2, 2	Mexico, U.S.
34	18-19	M	Student	English, Korean	1, 10	Germany, Korea
35	18-19	M	Student/Part-time employee	English, Spanish	2, 2	U.S., U.S.
36	25-29	M	Student	English	2	U.S.

System Version and Script Order

Subject #	Non-Contingent	Constant Rule-Based	Linear-Decayed Rule-Based	Script Order
1	2	3	1	2 ; 3 ; 1
2	3	1	2	2 ; 3 ; 1
3	3	1	2	1 ; 3 ; 2
4	2	3	1	2 ; 1 ; 3
5	3	1	2	1 ; 2 ; 3
6	2	3	1	1 ; 3 ; 2
7	3	1	2	2 ; 1 ; 3
8	1	2	3	3 ; 1 ; 2
9	2	3	1	1 ; 2 ; 3
10	1	2	3	2 ; 3 ; 1
11	1	2	3	1 ; 3 ; 2
12	1	2	3	1 ; 2 ; 3
13	3	2	1	1 ; 2 ; 3
14	3	2	1	2 ; 3 ; 1
15	1	3	2	1 ; 2 ; 3
16	3	2	1	1 ; 3 ; 2
17	1	3	2	1 ; 3 ; 2
18	3	2	1	2 ; 1 ; 3
19	1	2	3	2 ; 1 ; 3
20	3	1	2	3 ; 1 ; 2
21	2	1	3	1 ; 3 ; 2
22	2	1	3	1 ; 2 ; 3
23	1	3	2	2 ; 3 ; 1
24	2	1	3	2 ; 1 ; 3
25	2	3	1	3 ; 1 ; 2
26	2	1	3	3 ; 2 ; 1
27	3	2	1	3 ; 2 ; 1
28	2	1	3	2 ; 3 ; 1
29	1	3	2	2 ; 1 ; 3
30	1	3	2	3 ; 2 ; 1
31	2	1	3	3 ; 1 ; 2
32	1	2	3	3 ; 2 ; 1
33	2	3	1	3 ; 2 ; 1
34	3	1	2	3 ; 2 ; 1
35	3	2	1	3 ; 1 ; 2
36	1	3	2	3 ; 1 ; 2

Question #1 Ratings

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

Question #2 Ratings

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

Question #3 Ratings

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

Question #4 Ratings

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

Question #5 Ratings

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

Question #6 Ratings

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

Question #7 Ratings

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

Question #8 Ratings

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

Question #9 Ratings

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

Question #10 Ratings

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

Question #11 Ratings

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

Comparison-based Questionnaire Results

Subject #	Preference Order (Most, 2nd Most, Least)	Naturalness Judgment (Most, 2nd Most, Least)
1	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent
2	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based	Non-Contingent, Linear-Decayed Rule-Based, Constant Rule-Based
3	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent
4	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent
5	Constant Rule-Based, Non-Contingent, Linear-Decayed Rule-Based	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent
6	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based
7	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based	Non-Contingent, Linear-Decayed Rule-Based, Constant Rule-Based
8	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent
9	Constant Rule-Based, Non-Contingent, Linear-Decayed Rule-Based	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent
10	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent
11	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based
12	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based
13	Constant Rule-Based, Non-Contingent, Linear-Decayed Rule-Based	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent
14	Non-Contingent, Linear-Decayed Rule-Based, Constant Rule-Based	Non-Contingent, Linear-Decayed Rule-Based, Constant Rule-Based
15	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent
16	Constant Rule-Based, Non-Contingent, Linear-Decayed Rule-Based	Constant Rule-Based, Non-Contingent, Linear-Decayed Rule-Based
17	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent
18	Non-Contingent, Linear-Decayed Rule-Based, Constant Rule-Based	Non-Contingent, Linear-Decayed Rule-Based, Constant Rule-Based
19	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent
20	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent

Subject #	Preference Order (Most, 2nd Most, Least)	Naturalness Judgment (Most, 2nd Most, Least)
21	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based	Constant Rule-Based, Non-Contingent, Linear-Decayed Rule-Based
22	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based	Non-Contingent, Linear-Decayed Rule-Based, Constant Rule-Based
23	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based
24	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based
25	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent
26	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based
27	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based
28	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based
29	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent	Constant Rule-Based, Linear-Decayed Rule-Based, Non-Contingent
30	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based
31	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based	Non-Contingent, Constant Rule-Based, Linear-Decayed Rule-Based
32	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent
33	Constant Rule-Based, Non-Contingent, Linear-Decayed Rule-Based	Constant Rule-Based, Non-Contingent, Linear-Decayed Rule-Based
34	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based	Linear-Decayed Rule-Based, Constant Rule-Based, Non-Contingent
35	Non-Contingent, Linear-Decayed Rule-Based, Constant Rule-Based	Non-Contingent, Linear-Decayed Rule-Based, Constant Rule-Based
36	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based	Linear-Decayed Rule-Based, Non-Contingent, Constant Rule-Based

Appendix E: Analyses on the Data from the User Study

This appendix contains tables representing analysis data from the data collected during the user study. First correlation tables on the rating question result data are presented, along with supplementary tables describing key correlation values (such as those pertaining to the hypotheses). Correlations are present based on all rating question data, followed by rating question data for each dialog system version. This is followed by the p values from the one-tailed, paired-sample t tests done on the rating question results. After the p values for each rating question dimension, the means and p values are presented based on different combinations of rating question dimensions (p values are measured here using one-tailed, paired-sample t tests on averaged results of the rating question combinations). Finally, the chi-squared test results done on the comparison-based questionnaire results are presented.

Correlations Among Rating Questions Based on All Subjects and All Dialog System Versions

Q #	**Scale**	**Q #**	**Scale**	**Q #**	**Scale**	**Q #**	**Scale**
1	Emotional Rapport	2	Cognitive Rapport	3	Helpful	4	Trustworthy
5	Likeable	6	Natural	7	Enjoyable	8	Human-like
9	Persuasive	10	Preference	11	Recommendable		

Q #	1	2	3	4	5	6	7	8	9	10	11
1	1	0.82	0.72	0.7	0.71	0.73	0.66	0.59	0.67	0.44	0.77
2	0.82	1	0.69	0.65	0.68	0.74	0.59	0.56	0.58	0.29	0.66
3	0.72	0.69	1	0.85	0.79	0.64	0.69	0.51	0.7	0.27	0.66
4	0.7	0.65	0.85	1	0.86	0.68	0.77	0.57	0.79	0.25	0.67
5	0.71	0.68	0.79	0.86	1	0.72	0.78	0.7	0.79	0.33	0.65
6	0.73	0.74	0.64	0.68	0.72	1	0.68	0.73	0.68	0.34	0.66
7	0.66	0.59	0.69	0.77	0.78	0.68	1	0.66	0.76	0.44	0.68
8	0.59	0.56	0.51	0.57	0.7	0.73	0.66	1	0.72	0.46	0.65
9	0.67	0.58	0.7	0.79	0.79	0.68	0.76	0.72	1	0.46	0.82
10	0.44	0.29	0.27	0.25	0.33	0.34	0.44	0.46	0.46	1	0.49
11	0.77	0.66	0.66	0.67	0.65	0.66	0.68	0.65	0.82	0.49	1

This supplementary table describes correlation values for questions pertaining to hypotheses.

Natural ↔ Human-like Correlation	0.73
Likeable ↔ Enjoyable Correlation	0.78
Likeable ↔ Preference Correlation	0.33
Enjoyable ↔ Preference Correlation	0.44
Natural ↔ Preference Correlation	0.34
Human-like ↔ Preference Correlation	0.46

Correlations Among Rating Questions Based on All Subjects and the Non-Contingent Dialog System

Q #	**Scale**	**Q #**	**Scale**	**Q #**	**Scale**	**Q #**	**Scale**
1	Emotional Rapport	2	Cognitive Rapport	3	Helpful	4	Trustworthy
5	Likeable	6	Natural	7	Enjoyable	8	Human-like
9	Persuasive	10	Preference	11	Recommendable		

Q #	1	2	3	4	5	6	7	8	9	10	11
1	1	0.84	0.69	0.64	0.66	0.75	0.65	0.55	0.64	0.51	0.85
2	0.84	1	0.57	0.6	0.61	0.74	0.59	0.52	0.56	0.28	0.74
3	0.69	0.57	1	0.78	0.74	0.58	0.69	0.5	0.7	0.46	0.66
4	0.64	0.6	0.78	1	0.87	0.58	0.77	0.56	0.81	0.43	0.69
5	0.66	0.61	0.74	0.87	1	0.64	0.82	0.71	0.82	0.47	0.67
6	0.75	0.74	0.58	0.58	0.64	1	0.69	0.71	0.65	0.43	0.72
7	0.65	0.59	0.69	0.77	0.82	0.69	1	0.7	0.79	0.49	0.7
8	0.55	0.52	0.5	0.56	0.71	0.71	0.7	1	0.72	0.45	0.59
9	0.64	0.56	0.7	0.81	0.82	0.65	0.79	0.72	1	0.46	0.8
10	0.51	0.28	0.46	0.43	0.47	0.43	0.49	0.45	0.46	1	0.49
11	0.85	0.74	0.66	0.69	0.67	0.72	0.7	0.59	0.8	0.49	1

This supplementary table describes correlation values for questions pertaining to hypotheses.

Natural ↔ Human-like Correlation	0.71
Likeable ↔ Enjoyable Correlation	0.82
Likeable ↔ Preference Correlation	0.47
Enjoyable ↔ Preference Correlation	0.49
Natural ↔ Preference Correlation	0.43
Human-like ↔ Preference Correlation	0.45

Correlations Among Rating Questions Based on All Subjects and the Constant Rule-Based Dialog System

Q #	**Scale**	**Q #**	**Scale**	**Q #**	**Scale**	**Q #**	**Scale**
1	Emotional Rapport	2	Cognitive Rapport	3	Helpful	4	Trustworthy
5	Likeable	6	Natural	7	Enjoyable	8	Human-like
9	Persuasive	10	Preference	11	Recommendable		

Q #	1	2	3	4	5	6	7	8	9	10	11
1	1	0.84	0.76	0.72	0.76	0.79	0.73	0.68	0.71	0.53	0.74
2	0.84	1	0.78	0.67	0.73	0.82	0.61	0.63	0.6	0.37	0.69
3	0.76	0.78	1	0.87	0.82	0.75	0.68	0.52	0.71	0.24	0.69
4	0.72	0.67	0.87	1	0.85	0.74	0.82	0.55	0.81	0.3	0.73
5	0.76	0.73	0.82	0.85	1	0.8	0.79	0.7	0.8	0.38	0.71
6	0.79	0.82	0.75	0.74	0.8	1	0.72	0.74	0.66	0.35	0.63
7	0.73	0.61	0.68	0.82	0.79	0.72	1	0.69	0.79	0.48	0.74
8	0.68	0.63	0.52	0.55	0.7	0.74	0.69	1	0.67	0.47	0.66
9	0.71	0.6	0.71	0.81	0.8	0.66	0.79	0.67	1	0.57	0.85
10	0.53	0.37	0.24	0.3	0.38	0.35	0.48	0.47	0.57	1	0.55
11	0.74	0.69	0.69	0.73	0.71	0.63	0.74	0.66	0.85	0.55	1

This supplementary table describes correlation values for questions pertaining to hypotheses.

Natural ↔ Human-like Correlation	0.74
Likeable ↔ Enjoyable Correlation	0.79
Likeable ↔ Preference Correlation	0.38
Enjoyable ↔ Preference Correlation	0.48
Natural ↔ Preference Correlation	0.35
Human-like ↔ Preference Correlation	0.47

Correlations Among Rating Questions Based on All Subjects and the Linear-Decayed Rule-Based Dialog System

Q #	**Scale**	**Q #**	**Scale**	**Q #**	**Scale**	**Q #**	**Scale**
1	Emotional Rapport	2	Cognitive Rapport	3	Helpful	4	Trustworthy
5	Likeable	6	Natural	7	Enjoyable	8	Human-like
9	Persuasive	10	Preference	11	Recommendable		

Q #	1	2	3	4	5	6	7	8	9	10	11
1	1	0.81	0.74	0.76	0.71	0.65	0.58	0.52	0.67	0.27	0.7
2	0.81	1	0.76	0.72	0.75	0.67	0.6	0.53	0.6	0.23	0.53
3	0.74	0.76	1	0.87	0.82	0.58	0.72	0.51	0.69	0.14	0.62
4	0.76	0.72	0.87	1	0.89	0.7	0.73	0.6	0.76	0.05	0.61
5	0.71	0.75	0.82	0.89	1	0.72	0.74	0.7	0.75	0.14	0.56
6	0.65	0.67	0.58	0.7	0.72	1	0.65	0.74	0.74	0.23	0.62
7	0.58	0.6	0.72	0.73	0.74	0.65	1	0.57	0.69	0.35	0.59
8	0.52	0.53	0.51	0.6	0.7	0.74	0.57	1	0.79	0.44	0.69
9	0.67	0.6	0.69	0.76	0.75	0.74	0.69	0.79	1	0.29	0.79
10	0.27	0.23	0.14	0.05	0.14	0.23	0.35	0.44	0.29	1	0.42
11	0.7	0.53	0.62	0.61	0.56	0.62	0.59	0.69	0.79	0.42	1

This supplementary table describes correlation values for questions pertaining to hypotheses.

Natural ↔ Human-like Correlation	0.74
Likeable ↔ Enjoyable Correlation	0.74
Likeable ↔ Preference Correlation	0.14
Enjoyable ↔ Preference Correlation	0.35
Natural ↔ Preference Correlation	0.23
Human-like ↔ Preference Correlation	0.44

P-Values for Rating Question Results T-Tests

Q #	Scale	(Non-Contingent, Constant Rule-Based) P-Value	(Non-Contingent, Linear-Decayed Rule-Based) P-Value	(Constant Rule-Based, Linear-Decayed Rule-Based) P-Value
1	Emotional Rapport	0.3602	0.2152	0.2855
2	Cognitive Rapport	0.2929	0.2974	0.1043
3	Helpful	0.3269	0.2213	0.3720
4	Trustworthy	0.3684	0.4420	0.2986
5	Likeable	***0.0575***	0.2614	***0.0058***
6	**Natural**	0.3000	0.1551	0.4173
7	Enjoyable	***0.0622***	0.3269	***0.0203***
8	**Human-like**	0.3887	***0.0397***	***0.0625***
9	Persuasive	0.4444	0.1014	0.1269
10	**Preference**	0.2065	***0.0052***	***0.0143***
11	Recommendable	0.3503	0.1621	0.1164

Rating Question Combinations Results

Means for the Rating Question Combinations

Rating Questions Averaged	Non-Contingent	Rule-Based	Linear
Likeable/Enjoyable/Preference	4.48	4.35	4.74
Likeable/Enjoyable	5.26	4.99	5.38
Likeable/Preference	4.1	4.04	4.44
Enjoyable/Preference	4.08	4.03	4.4
Natural/Human-Like	3.99	4.07	4.25
Natural/Preference	3.57	3.71	3.93
Human-Like/Preference	3.33	3.44	3.79
Natural/Human-Like/Preference	3.63	3.74	3.99

T-Test Results (P-Values) for the Rating Question Combinations

Rating Questions Averaged	Non-Contingent/Rule-Based	Non-Contingent/Linear	Rule-Based/Linear
Likeable/Enjoyable/Preference	0.1580	***0.0362***	***0.0040***
Likeable/Enjoyable	***0.0383***	0.2711	***0.0075***
Likeable/Preference	0.3325	***0.0116***	***0.0040***
Enjoyable/Preference	0.3568	***0.0165***	***0.0060***
Natural/Human-Like	0.3177	***0.0434***	0.1914
Natural/Preference	0.2140	***0.0102***	0.1231
Human-Like/Preference	0.2203	***0.0015***	***0.0140***
Natural/Human-Like/Preference	0.2305	***0.0045***	***0.0767***

Chi-Squared Test Values for Comparison-based Questionnaire Results

	(Non-Contingent, Constant Rule-Based)	(Non-Contingent, Linear-Decayed Rule-Based)	(Rule-Based, Linear-Decayed Rule-Based)
Most Preferred	0.7728	0.6831	0.7728
2nd Most Preferred	0.6831	0.5186	0.5186
Least Preferred	0.7728	0.7728	0.6831
Most Natural	0.7728	0.7728	0.6831
2nd Most Natural	0.1967	0.1967	0.4142
Least Natural	0.2340	0.1489	0.3613
Preferred Results	0.7389	0.5050	0.5050
Natural Results	0.7389	0.3173	0.7389

Chi-squared Test Scores for Each Question on the Comparison-based Questionnaire

Chi-Test Scores	Most Preferred	2nd Most Preferred	Least Preferred	Most Natural	2nd Most Natural	Least Natural
	0.9200	0.7788	0.9200	0.9200	0.3679	0.3385

Vita

Michael Durcholz was born in Landstuhl, Germany on May 31, 1987. He has received the Gates Millennium Scholarship throughout his collegiate career, both undergraduate and graduate levels. Michael started his undergraduate coursework at The University of Texas at El Paso in 2005. He transferred to The University of Texas at Austin and completed his Bachelor's degree in computer science in 2009. While there, Michael conducted independent research related to computer vision for weather forecasting purposes. Throughout his collegiate career, Michael worked for numerous organizations, including the National Aeronautics and Space Administration, the Texas Department of Transportation, and the Army Research Laboratory.

Michael then pursued a Master's degree back at The University of Texas at El Paso. During his graduate studies, Michael was under the direction of Dr. Nigel Ward. He became a member of the Interactive Systems Group, a research group focused on human-computer interaction, where he learned more about computational linguistics and its application to vocal communication. A paper based on his thesis work is currently submitted to an international speech communication conference (Interspeech). In addition, Michael published a paper with Dr. Vladik Kreinovich on developing a generic notion of genericity.

For the future, Michael will be working full-time for ExxonMobil in Houston starting in July 2012. However, he has a long-term goal of pursuing a PhD within either human-computer interaction or computer graphics/vision.

Permanent address: 10541 Kendall Street

El Paso, Texas 79924

This thesis was typed by Michael Durcholz.