2009-01-01

Appropriate Prosodic Variation is Valued by Users

Rafael Escalante-Ruiz
University of Texas at El Paso, rafa.escalante@gmail.com

Appropriate Prosodic Variation is Valued by Users

RAFAEL ESCALANTE-RUIZ

Department of Computer Science

APPROVED:

Nigel G. Ward, Chair, Ph.D.

David G. Novick, Ph.D.

Jon Amastae, Ph.D.

Patricia D. Witherspoon, Ph.D.

Dean of the Graduate School

to my

MOTHER, FATHER and SISTER

with love

Appropriate Prosodic Variation is Valued by Users

by

RAFAEL ESCALANTE-RUIZ

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

August 2009

Table of Contents

Appendix

List of Figures

List of Tables

Abstract

When humans converse they can detect and respond to their interlocutor's fleeting emotional changes. This ability is especially important in tutoring situations, because effective tutors assess the learner's need for help and encouragement and act appropriately at the correct times. The most effective tutors are able, in addition, to detect emotions in the learner such as uncertainty, over-confidence, and enthusiasm, and then react with emotionally appropriate behaviors.

Current computer speech-based systems lack the ability to detect the user's emotional changes and thus are unable to respond in an emotionally appropriate way. While there has been research in emotion recognition and generation in tutoring systems, it has focused mainly on words. Previous research by Hollingsed and Ward [7] showed that varying the words of the acknowledgments of a speech-based tutoring system made the system better liked by users [7, 13]; however, the users complained about the fixed prosody of the system's acknowledgments.

The present research aimed to create a rule-based model of the behavior and response prosody of a human tutor and to test its effectiveness. The first goal was to discover the different types of emotions expressed by the tutor; to do this, 339 tutor acknowledgments contained in 16 tutor-student interactions were analyzed by a group of raters to find common emotion types. The second goal was to discover how these emotions were expressed, assuming that the cues to interpretation correspond to the speaker's intent. To do this, responses rated high for each emotion type were grouped and analyzed to find commonalities in their prosody. The third goal was to find when the tutor used each emotion. To do this, the conversations were again analyzed to find commonalities in the context-of-occurrence of each emotion. Each commonality was quantified and the result was a set of rules that described the behavior of the tutor. For example, talking with a warm tone of voice when the student is uncertain, responding authoritatively to

keep control when the student is over-confident, or responding with high energy when the student is enthusiastic. To test the effectiveness of these rules, they were integrated into a Wizard-of-Oz [8] tutoring system and 21 students were asked to interact with it and compared it to a system that produced random acknowledgments. The student's perceptions of friendliness and naturalness were higher when using the rule-based tutoring system. Although only one of the three measures were significant, the users tended to prefer the system that was appropriately emotionally responsive. This suggests that emotional modeling with prosodic variation can be effective.

Chapter 1

Introduction

Giving appropriate feedback to students in intelligent tutoring systems has been a topic of much research. Tutorial systems have practical value for helping students memorize or review subjects such as multiplication tables, standard abbreviations, famous people, and dates [5].The importance of positive feedback has been discussed by Fossati [3], in the context of a text-based tutorial system. However, the question of which feedback to givene is also important. Many researchers [9, 10, 1, 2] have suggested that for this, as for other problems, it is necessary to diagnose and act on the student's affective state. Porayska-Pomsta [9, 10], for example, indicated that three factors—student confidence, interest and effort—were important. Others [1, 2] have shown that these states can be inferred from the context in the interaction in a text-based system.

To better infer the student's affective state, we can use speech-based tutorial systems since the speech signal brings additional information about the student's state. Many researchers [] have discussed the audio correlates of different emotions. However, this has been applied in tutorial dialog systems only by Tsukahara and Ward [14] and Hollingsed [7], who showed that varying the words used to acknowledge correct answers, based on information about the user inferred from his prosody, made the system better-liked by the users. However, the prosody of the responses was always the same, and the users commented that this was sometimes inappropriate. The prosody of back-channels and acknowledgments have received attention [13, 4, 11]; however, the specific aspects of acknowledgment prosody used in tutorial contexts has not been studied, nor has the actual value of manipulating acknowledgment prosody in any context. Therefore, I decided to vary the prosody of the system's responses. I hypothesized that this would make the

system seem more natural and warm.

1.1 depicts a conceptual diagram of the system.

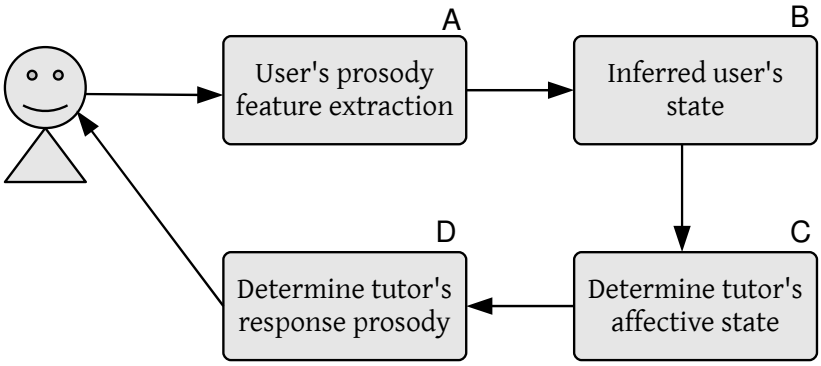

Figure 1.1: Conceptual diagram of the system.

1.1 Hypotheses

The following hypotheses lay the foundation for this thesis.

1. Students will prefer a tutor with rule-based prosody production over a tutor with random prosody.

2. Students will find the rule-based tutor more natural and friendly than a tutor with random prosody.

   The first hypothesis predicts that the the response to the question "Which system would you prefer to use?" after the subject has interacted with a system will favor the rule-based system. The second hypothesis predicts that the responses to the questions "How natural was each system?" and "How friendly was each system" will be higher for the tutor with rule-based prosody.

Chapter 2

Methodology

In this chapter, I turn to the experimental methods I used to test the hypotheses. My approach involved three main steps: Developing rules for varying prosody in the tutor's feedback utterances, building a system embodying these rules, and testing the system with a population of human subjects. The two hypotheses were tested by applying a post-experiment questionnaire.

2.1 Developing Rules

The first step of my experimental approach involved developing prosody rules for the tutor. To do this, I analyzed the prosody of responses of a human tutor in the domain of memory quiz tasks. Each task consisted of one prompt, for example, "Name any eight I-10 exits in order" and a set of answers.

The following section describes the corpus and the various analysis methods tried, including those which were later superseded by better analysis.

2.1.1 Corpus Collection

The corpus, collected by Tasha Hollingsed, consists of 16 conversations, each between two introductory computer science students, one in the role of a tutor, and the other the role of a student.

The student in the role of a tutor was the same in all conversations, and was selected upon the criteria of friendliness, flexibility and being generally interesting to listen to

from a set of six tutors [6]. The tutor-selection process had each of the six possible tutors interact with a student, and the tutor was given a list of items that the student would have to say in a particular order. Each possible tutor was instructed to give hints and reveal the answers when they deemed necessary. The reason for modeling a single tutor for the entire corpus was that the variety of responses differed for each tutor (for example, some tutors only used "uh-huh" or "okay" in response) [6]. The selected tutor was friendly and upbeat during the interactions with student and remained positive regardless of the students' performance. When the questions were hard to answer by the students, she gave hints frequently and gave away answers to prevent students from becoming too frustrated. Her acknowledgment choice was atypical in the choice and variety of responses, but this seemed to be part of the style that made her effective.

After the tutor was selected, she was paired with several of her peers in the same introductory course using four of the same quizzes that were used in the tutor selection phase. However, after several interactions, one quiz was deemed inappropriate because it contained material that was unfamiliar to students. Data for the remaining three quizzes was retained in the corpus.

Each utterance in the corpus was labeled using Didi [12] with a pre-established set of tags, such as "guess acknowledgment," "correct guess," and "incorrect guess," to facilitate subsequent analysis of the dialog acts. The final dialog transcriptions include the transcribed utterances, the tags, and the utterance lengths all for both channels in each audio file.

2.1.2 Initial Examination of the Tutor Responses

I sought to develop rules that would explain the differences in prosody in the tutor's responses to the learner. Doing this required finding instances where the tutor used different prosody for identical lexemes, as differences in such cases should be attributable to different meanings that the tutor intended to express. Likewise, developing the rules also

required finding instances where the tutor used the same prosody for different lexemes, as these cases should show a common affective meaning despite the differences in words.

Therefore, I reviewed, in its dialog context, every instance of the tutor's feedback responses to identify these differences and similarities. First, I grouped the responses by lexical item (e.g., "very good," "good job," "uh-huh," "mm-hmm") and annotated each response with my perception of the response's prosodic function. I then reviewed my notes and generalized these for the most common responses ("very good," "good job," and "uh-huh," which comprised 62.5% of the corpus) into five dimensions of emotion conveyed by prosody: *energy*, *warmth*, *empathy*, *condescension*, and *impatience*. This was my initial attempt to identify the relevant categories in box C of Figure 1.1.

I was also interested in the tutor's dynamic assessment of the student, so for each of the tutor's responses I recorded my subjective impression of the tutor's assessment. Examples of these annotations include "the tutor was already confident that the student would get this answer right," "the tutor feels the student is doing great but she is tired," and "the tutor thinks the student is doing very bad and wants to make him feel good."

2.1.3 Context-Free Perceptions of the Responses

Another way of looking at the tutor's responses is to do so without knowledge of the responses' dialog context. It is possible that a response's dialog function could be identified by the response's words and prosody alone, without the context of the student's utterance to which the tutor responded or, for that matter, any of the previous student-tutor exchanges. Accordingly, I reviewed each of the tutor's responses without the responses' dialog context. I grouped the responses by lexical item ("very good," "good job," and "uh-huh"), and listened to responses within each lexical group to identify the variations in prosody.

Using the five emotional dimensions previously identified, I rated the strength of each response on a scale from 1 to 7, where 1 meant weakly present and 7 meant strongly

present. To assess the reliability of my coding, I asked five members of the UTEP's Computer Science Interactive Systems Group to classify 20 of the responses (out of 212 responses for the selected lexical items) that had been perceived in various, sometimes contradictory, ways in my initial subjective ratings. The members were not blind to the purpose of this assessment. Analysis of our coding with Fleiss's Kappa indicated that only the codings for *energy* and *warmth* were reliable.

2.1.4 Initial Development of Rules

As the next step towards developing predictive rules, I examined the prosodic characteristics of the responses associated with each of the emotional dialog functions, aiming to uncover the mapping between boxes C and D of Figure 1.1. For each dialog function, I analyzed the responses that were rated high for that function and looked for common prosodic features. For responses rated high for *warmth*, I observed that the ending syllables were elongated and some had the creaky voice, which is a glottalization. Responses rated high for *energy* had higher volume and sharper pitch slopes. Responses rated high for *condescension* had elongation of syllables early in utterance, variation in pitch (i.e., pitch was not flat), and variation in volume. Responses rated high for *impatience* were noticeably shorter in length and contained a pitch upturn at the end. Responses rated high for *empathy* were elongated and some had vibrato. Table 2.1 summarizes the common prosodic features.

2.1.5 Differences in Perception of Responses when Listened to without and with Context

After evaluating the responses out of context, I evaluated them again, this time listening to the entire tutoring session. Surprisingly, almost all of the out-of-context "hunches" regarding the emotions conveyed were confirmed when listening to the responses in context. For example, in Table 2.2, the "good job[2]" response shows similar perception when listened to in context.

Table 2.1: Emotional dimensions and their common prosodic features

Perceived feeling	Prosodic features
Energy	Volume higher than average, sharp pitch slopes, noticeable pitch modulation
Warmth	Elongated ending syllable, creaky voice
Empathy	Pitch is not flat, vibrato, elongated syllables
Condescension	Elongated starting syllable, longer utterance, pitch variation
Impatience	Short length, pitch upturn

One exception to these matches happened when the response signaled the end of the conversation. The response included an upturn, sometimes with atypical pitch variation that was perceived as condescending in the out-of-context evaluation. When heard in context, it simply shows that the tutor was ending the tutoring session and was pleased.

I next worked on validating my preliminary observations of in- and out-of-context responses with external raters. The members of the research group listened to the responses with the lowest Kappa in the previous group evaluation. The tutor's responses were played in random order; all the responses played in context (at least ten seconds of the conversation before and after the response) and then all the responses were then played again out of context. The group members rated each response as either *empathetic, warm, energetic, helpful, enthusiastic, impatient, condescending, anxious, absent-minded, or other.* After hearing each response, the group briefly compared their judgments. The Fleiss Kappa test for agreement among multiple raters indicated moderate agreement ($K = 0.43$). Although the raters' judgments were generally in agreement with respect to individual responses, their judgments for the in-context responses did not always match their judgments for the same responses out of context, as illustrated in Table 2.3. For example, some responses that were rated as condescending and impatient

Table 2.2: Examples of perception changes when responses are listened to in- and out-of-context

Response	Out-of-context perception	In-context perception
Good job[1]	Not very empathetic, possibly tired, a bit condescending.	The intonation says "this was a hard question, but you got it right".
Good job[2]	Empathetic and warm. The tutor is calm, waiting patiently for next response. No discernible condescension.	The tutor is satisfied with the response, thinks it is a good start and expects a correct response afterward.
Good job[3]	The tutor is impatient, although she thinks that the student is doing well. She is not very warm, but at the same time she is pleased with the student's performance.	The student was facing difficulties before this question, but he has got these two right, so she doesn't want to interrupt the stream of good answers (the student is quickly giving responses).
Uh-huh, good job	Energetic but condescending, the student was doing badly or is unsure about the guess.	Tutor marks end of section. The tutor is pleased about the student's performance.
Very good[1]	Empathetic and calm response. It could also be considered condescending. The student probably was facing difficulties responding to this question, or the student was failing in general.	The tutor wants to give confidence to the student, who seems insecure in her answers. Upturn in student's guess.
Very good[2]	Cold response; pitch is flat. The tutor is receiving and expecting correct answers at a comfortable pace. The tutor does not need to encourage the student.	The tutor is responding flatly because the student is giving correct responses after only a moment of difficulty.

Table 2.3: Perception changes of emotions when listened in context.

Out-of-context perception	In-context perception
Condescension	Little or no condescension, warmth
Energetic for no reason	Coming towards the end
Indifferent back-channel	"There's something missing in your answer"
Indifferent back-channel	Momentarily unsure about the correctness of the answer

when heard out of context were judged as energetic, warm and enthusiastic when heard in context. Some responses that were relatively featureless, such as back-channels, were rated inconsistently when heard out of context, but judged helpful, warm and empathetic when heard in the context of the conversation. This result was unexpected and may have been attributable to the empathetic nature of back-channels.

2.1.6 Finding the Dialog Functions Served by Each Prosodic Feature

After these preliminary observations, I aimed to find the dialog functions that each prosodic feature seen in the acknowledgments served. I created a set of tentative rules that linked the context and feeling of the student with the prosody of the tutor's responses by directed listening rather than systematic analysis. Table 2.4 shows the tentative rules created after the preliminary observations. The accuracy of these rules was not tested, as they were only the basis for the development of the quantified rules.

2.2 Technical Implementation

This section describes how the system that computed the prosody and chose the correct prosody for each acknowledgment was built.

The tutoring system consisted of three key parts: the prosody classifier, the rules

Table 2.4: Tentative rules.

Context & Feeling of Student	Tutor's Prosody	Meaning
Question was hard	Flat pitch, lengthened, lengthened first syllable	Praise
Response is satisfactory and expected continued good performance	Creaky, elongated vowels	Satisfaction
Getting back on track after difficulties	Shorter utterance lengths, pitch upturn	Satisfaction, avoiding interruption
Low confidence, delay in response	Pitch is not flat, vibrato, elongated syllables	Perfunctory encouragement
Student is performing poorly	Volume is higher than average, elongated vowel in starting syllable, longer utterances, pitch vibrato	Reassurance
Student is giving correct answers in quick succession	Flat pitch, may have a creaky voice or elongated vowel in ending syllable	Encouragement

for determining which prosodic pattern are used in each situation, and the response production subsystem. These components are shown in Figure 2.1.

2.2.1 Classifying the Student's Pitch Slope

The prosody classifier consisted of three prosody contour classifiers that worked independently. Each of these contour classifiers analyzed the utterance's pitch to determine whether an upturn, downslope, or flat contour was present. The final prosody classifier was built using the output of the contour classifiers to determine the prosody of an utterance. The contour classifiers were originally part of the Yesman [14] program.

After developing the prosody classifier, I evaluated its performance, because the final

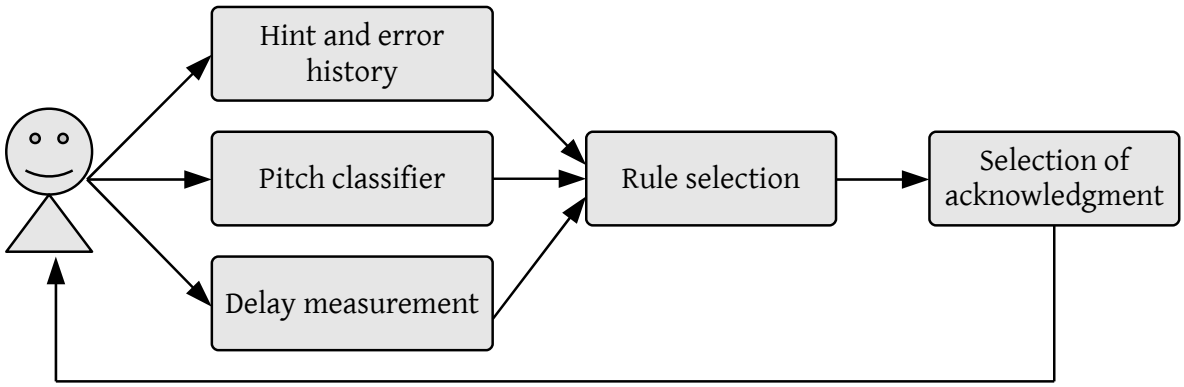

Figure 2.1: System implementation. The rule selection module implements a direct mapping from A to D in Figure 1.1.

rules would depend on the prosody classification to determine how to respond. I measured the accuracy and coverage of the classifications by comparing the result of the prosody classifier to evaluations by an external rater, who used the Didi program to view and hear the examples, and classify and label the prosody of the tutor's responses as either containing a downslope, an upturn or a flat prosody. The rater was instructed to look at the pitch contour in the Didi display and to hear the perceived prosody. If there was a discrepancy between the displayed and heard prosody, priority was given to the heard prosody. The coverage and accuracy of the rules was measured by the formulas

$$\text{Accuracy} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

$$\text{Aoverage} = \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}}$$

In these formulas, true positives were the number of responses in the corpus that agreed with the prosody detector (e.g., the prosody detector found a upturn and there was an upturn in the manually evaluated response); false positives were responses that were falsely identified by the prosody detector (e.g., the prosody detector found a downturn when there was none); true negatives were the number of responses which matched the prosody detector's assessment of a lack of contour (e.g., the prosody detector determined

that there was no upturn in the response and there was no upturn in the manually evaluated response); and false negatives were responses that contained a prosodic contour that was undetected by the prosody detector (e.g., the response contained an upturn but the prosody detector did not detect it).

A preliminary evaluation after the first build indicated that all three classifiers performed poorly. For example, the accuracy of the downslope classifier was 0.23 and the coverage was 0.37. While perfect accuracy and coverage is not unreasonable to expect, this seemed too low to be usable in the system. To improve accuracy and coverage, various systematic methods are known, but for simplicity, I iteratively adjusted the classification parameters by reviewing the instances where the classifiers failed. Small modifications to the parameters cumulatively produced large changes in accuracy, as shown in Table 2.5. The *minimum downslope length* (MDL) parameter defines the minimum utterance length required to be considered a downslope; downslopes shorter than the MDL will not be counted. The *minimum fall rate* (MFR) parameter defines the minimum percentage of pitch fall that a downslope should exhibit to be considered a downslope. As might be expected, reducing the MDL and MFR tended to improve the coverage but decrease the accuracy. In some cases, both accuracy and coverage improved, as shown in Table 2.5.

2.2.2 Combining the Output of the Pitch Classifiers into a Single Pitch Determination

This section describes how the multiple outputs of the pitch classifiers described in the previous section were consolidated into a single output that could be easily incorporated into the final rule set.

The purpose of analyzing student pitch slopes is to enable the application of rules to determine the prosody of the system's utterance. To simplify these rules, the detected pitch slope of a student's utterance was assumed to be either an upturn, a downslope or a flat. The preliminary pitch-slope detection algorithm could assign multiple slope values

Table 2.5: History of manual adjustments to the downslope classifier parameters.

Accuracy	Coverage	Manual adjustments
0.23	0.37	Initial
0.27	0.55	Decreased *minimum downslope length* to 60ms from 70ms
0.29	0.53	Decreased *minimum fall rate* to 0.008
0.25	0.55	Decreased *minimum fall rate* to 0.003
0.33	0.53	Increased *minimum fall rate* to 0.010
0.33	0.50	Increased *minimum fall rate* to 0.012
0.25	0.58	Decreased *minimum downslope length* to 40ms from 60ms

to a single utterance because the three pitch slope classifiers operated independently. Thus if an utterance contained both flat prosody at the middle of an utterance and downslope prosody towards the end, the utterance's prosody was likely to be classified as a downslope by the downslope classifier and as flat by the flat prosody classifier. The prosody production rules depend on the classification of the user's pitch as either an upturn, a downslope, or a flat. Originally, the classification was based on the last half of the utterance, but this accounted for a large part of the multiple-classifier misses. As a remedy, I tried reducing the scope of the classifiers to different last fractions of the utterance, from the original last half of the utterance down to the last tenth of the utterance. Limiting the scope of the classifiers to the last quarter of the utterance yielded the best results, decreasing the false positives of the flat pitch classifier.

Thus the prosody production rules of the system depend on a unique pitch determination per utterance, so the output of the independent pitch classifiers had to be combined into a single prosody classification.

2.2.3 Quantified Rules

The rules listed on Table 2.6 are the final set of rules that determine the tutor's response prosody productions, as they respond to the inferred context and feeling of the student. This table is a refinement of Table 2.4 and shows the rules as they were actually operational in the Wizard-of-Oz system; some had to qbe rewritten to refer to quantifiable factors (e.g., time since the student answered correctly or time since last response). In each rule, all conditions must be true for the rule to apply. Rules are checked in order. The conversion of the observations into rules was not independently verified, other than by the experiments themselves.

Specifically, the features referred to in the table were computed as follows:

- Delay from the onset of the tutor's acknowledgment of the previous correct answer to the onset of the current answer.
- Final pitch contour: flattish, downslope, upturn, measured using a linear approximation to the pitch contour over the last quarter of the guess. Upturn if rising at a rate of $>$ 50% per 100ms, downturn if $<$ -10% per 100ms over the last quarter.
- Number of hints before correct answer.
- Number of incorrect guesses before correct answer.
- Total incorrect guesses in the dialog so far.

The "tutor prosody" referred to in column 2 of the table was accomplished by applying the following manipulations to a neutral-sounding "good job" taken from the corpus:

- Elongation: using Audacity, addition of five pitch periods during the vowel of *job*
- Creaky: using Praat, altering the pitch throughout to superimpose a sawtooth pattern.
- Vibrato: using SOX with the –vibrato parameter using a period of 10Hz.

Table 2.6: Final rules. The letters in parenthesis show the correspondence to Figure 1.1.

Tutor feeling (C)	Tutor Prosody (D)	Student Feeling (B)	Conditions (A)
Warmth & praising	Elongated	Question is hard, possibly wanting praise	Delay $>$ 4 sec., hints $>$ 1
Praising & encouraging	Elongated and high energy (breathy, greater pitch rhange)	Not doing well, possibly discouraged	$\geq$ 2 incorrect guesses, $\geq$ 4 total incorrect guesses over the dialog
Keeping control	Creaky voice	Doing well, possibly feeling dominant	No incorrect guesses, pitch downslope
Welcoming a speed-up of pace	Upturn	Was not doing well, but now getting back on track	No hints, no incorrect guesses, $\geq$ 4 total incorrect guesses over the dialog
Reassuring	Vibrato	Low in confidence	Delay $>$ 2 seconds, pitch upturn
Expecting good performance	Creaky	Certain but still needing time to recall	Delay $>$ 3 sec.
No time to acknowledge	Acknowledgment omitted	Certain	Delay $<$ 2 sec., no incorrect guesses
Expecting continuation of good performance, at a slower pace	Creaky & elongated	Confident, but still needing time to recall	Delay $<$ 4 sec., no incorrect guesses
Neutral	Neutral	Neutral	Default

- High-energy: using an appropriate token taken from the corpus
- Upturn: using Praat, altering the pitch of the last voiced region.
- Creaky-elongated: adding creaky voice as described above to the elongated token.

Chapter 3

Experiments

This section describes the experiments I did to determine the effectiveness of the rule-based prosody production system. I compared the rule-based prosody production system to a system that randomly chose the response prosody.

3.1 Yesman

The spoken-dialog system used in the experiments was Yesman [14], a Wizard-of-Oz [8] (WoZ) system that computes prosodic features in user utterances and keeps track of the dialog state.

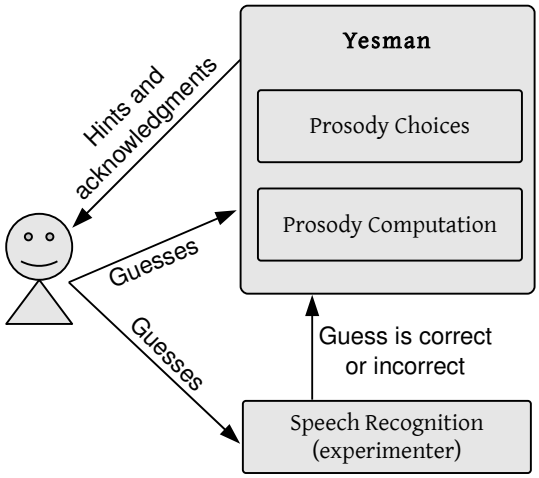

Figure 3.1: Experiment setup.

3.2 Participants

Twenty-six Computer Science undergraduate students participated in the experiments, of which I excluded four because the experiment conditions changed, and one because the audio recorder did not record the dialog. I changed the experiment protocol after the first four subjects mentioned that they rated the system that they used second higher than the one they used first because they felt more familiar with the experiment procedure after interacting with the first[1] system. Although this problem could not sway the experiment results because I swapped the system order for each participant, I wanted participants to be comfortable with the first system. To do this, I played back an audio clip of an example interaction with the system as part of the initial explanation, in order to let them hear how the system behaved and what they would be expected to do, and clarified that the experiment did not test their memorization skills. When I administered the questionnaire, I instructed the participants to focus on the systems' responses rather than in their own performance when answering the questions. After these changes, all participants acknowledged that they knew what to expect from the system. In the post-experiment interview all the subjects agreed that they had rated the system by the quality of the generated responses, without taking their own performance into account in the evaluation.

3.3 Experimental Method

The subject interacted with the rule-based prosody production system and the random prosody production system, following an IRB-approved procedure, as this section explains.

The subjects were first welcomed and asked to sign the consent form (see Appendix

[1]The "first" and "second" systems are the systems with which the subjects interacted the first time and the second time, respectively. The random and rule-based systems were chosen at random; if the first system happened to be the random system, the second one would be the rule-based system, and vice versa.

B). Afterward, the conditions of the experiment were explained to the participants: the subjects were told that the experiments involved interacting with a computer-based tutoring system, that they would have to memorize two lists of ten United States presidents, and that the conversations would be recorded. The subjects were not told the experiment objective, except that their memorization skills were not being tested. Both lists contained the names of ten presidents of the United States (see Appendix C); one consisted of the first ten presidents and the other one began with Jimmy Carter and was in reverse chronological order (see Appendix D). Both lists included a brief fact about the president to help memorization and was also the first hint that the system would give to the student when asked.

The system that they would test on first was randomly selected. Afterward, the participants were given a minute to memorize the first list of presidents (selected at random). The participants then interacted with the previously randomly selected system. This was the "first system," chronologically speaking. All conversations were recorded. An excerpt of a dialog with the system appears in Figure 3.2.

Afterward, the second list was given to the participant and given another minute to memorize the list. The first part of a questionnaire[2] was given. This questionnaire asked about the preferred system and friendliness of the system.

The next part of the experiment consisted of playing back the recordings of the entire conversations with the system to the subjects, so they could better evaluate the system without being under the pressure of memorizing and repeating items. After the participants heard both conversations, they were instructed to finish the other part of the questionnaire. After the questionnaire was filled out, I conducted a brief semi-structured interview to learn the participants' perceptions of the experiment. This had three parts. First, the participant was asked the question "which system would you prefer to use? A or B?" after interacting with the systems; second, the same question was asked after they heard their conversations with the systems. Third, if the participants changed their

[2]See Appendix A.

[System] In reverse chronological order, name ten presidents of the United States starting with Jimmy Carter.

[Subject] *Carter, Nixon.*

[System] No, that's not it.

[Subject] *Kennedy?*

[System] No, that's not it.

[Subject] *Hint.*

[System] His wife's name is Betty.

[Subject] *Ford.*

[System] Good job.

Figure 3.2: Transcript of a subject's interaction with the system

selection after hearing the recordings, I asked them why they had changed their selection. I annotated the responses that I received in Table 5.2 for the subjects who preferred the rule-based and in Table 5.1 for the subjects who preferred the random system.

To communicate with the Wizard-of-Oz system, each participant wore a headset (integrated headphones and microphone). This headset was connected to a preamplifier system to convert the low impedance signal of its microphone to a line level signal appropriate for splitting to the recorder and WoZ system. I did not use a dedicated preamplifier circuit; instead, I used the sound card of a laptop in loop-back mode to amplify the signal of the headset's microphone. The amplified voice signal was split and sent to the WoZ system and the audio recorder simultaneously. The productions of the WoZ system were also split and sent to both the participant's headset and the audio recorder. The conversation was recorded in two separate tracks in the audio recorder; the right channel recorded the user utterances and the left channel recorded the system's productions. The experimenter monitored the user utterances through headphones connected to the live

monitor output of the portable recorder.

The student and the experimenter were in the same room but they could not see each other; they were separated by a partition.

Chapter 4

Results

This chapter describes the results obtained from the experiments. The results analyze the subjects' average rating of the system responses and the difference between the preferences in pre- and post-listening questionnaires.

There were three independent variables: The system's response type: rule or random based; the quiz set used, set A or B; and the order in which the systems were presented to the subjects.

4.1 System Ratings

Every subject rated the systems in terms of naturalness and stated which system they preferred to use. They also re-rated the systems in terms of friendliness after listening to the recordings of their conversations with the system. Naturalness and friendliness ratings were in the scale of 1 to 7, where 1 was terrible and 7 excellent.

The questions about system preference and naturalness was administered immediately after the subject had interacted with both the rule and random-based systems. The results indicate that there was a perception of higher naturalness with the rule-based system immediately after the interaction. This result is significant by the t-test ($p < 0.05$). The subject's naturalness ratings after hearing their complete interaction for the systems also tended to be towards the rule-based system, although this preference was not statistically significant (see Table 4.1).

The questions of system preference, naturalness and friendliness were administered after the subjects had listened the recordings of their conversations with the system.

The question of friendliness was not administered in the pre-listening condition because I did not believe the subject could discern both naturalness and friendliness after having just interacted with the system under the pressure of answering correctly. Preliminary experiments showed that subjects were always concerned with their recall ability even when told that their performance was inconsequential, sometimes taking a long time trying to recall a forgotten item instead of asking for a hint. This observation led me to believe that their focused effort precluded a clear differentation between naturalness and friendliness when judging the systems.

Table 4.1: Preferences

System	Naturalness Before Listening	Naturalness After Listening	Friendliness After Listening
Rule	5.6 ($\sigma = 1.1$)	5.6 ($\sigma = 1.5$)	6.0 ($\sigma = 1.4$)
Random	5.0 ($\sigma = 1.5$)	5.3 ($\sigma = 1.3$)	5.7 ($\sigma = 1.3$)
Significance	$p < 0.05$	$p \approx 0.27$	$p \approx 0.24$

Chapter 5

Discussion

The thesis aimed to test the following hypotheses:

- That the participants would prefer the rule-based system over the random system.
- That the participants would find the rule-based system more natural and friendly than the random system.

Overall, the first hypothesis was not supported, although the tendency was in the direction expected. The second hypothesis was partly supported.

In this section I discuss possible reasons why the results came out the way they did.

There were marked differences in the responses of the post-experiment questionnaire of the subjects who preferred the random system. Subjects who preferred the rule-based system were more likely to observe differences in the prosody of the responses (three of the eight subjects who preferred the random system could not clearly identify a change of prosody between the systems. Four of the subjects who preferred the random system clearly perceived the random system as more natural, even when they did not consciously perceive a prosody difference. Only one of the subjects who preferred the random system noted a clear difference between the prosody variation of the systems).

Based on these comments from the subjects, it seems that the following factors, in addition to those we expected, influenced their perceptions of the two systems.

- Many subjects simply failed to perceive (at least consciously) any difference between the two systems.

- Some seemed to prefer the higher prosody variation offered by the random system, considering that to sound more natural, more "high energy", and to consider less variation to be "robotic".

- Some seemed to dislike variation, preferring instead more consistent, predictable responses.

- Several were swayed by differences in their own level of performance on the task, evidenced by the post-experiment interview.

- One subject seemed to perceive one of the acknowledgments as "machine"-like, possibly being sensitive to some details of the prosodic manipulations used to generate the various tokens.

- One may have been a affected by apparent system recognition errors when using one of the systems (the experimenter pressed the wrong keys when evaluating the subject's guess).

- One perceived a difference in delay, although it is not clear whether such a difference actually existed.

The fact that the preferences were weaker, not stronger, after the subjects listened to the recordings of their own interactions was a surprise. In previous research [7], the opposite was found.

One possible reason is in the nature of the instructions given the subjects. I revealed to them that the two systems they had interacted with were in fact following different rules. We then requested them to judge the two systems, and clarified that we were not interested in evaluation of their own memory performance on the quiz.

It is possible that some subjects misinterpreted this instruction to mean that they should ignore the context of the responses and the flow of the dialog, and these subjects may have just thought about the naturalness of the acknowledges as acoustic objects.

Table 5.1: Responses by Subjects who Preferred the Random System

Subject	Reason for Preference
105	"The systems sounded the same, although the random system waited before answering 'good job'."
114	"The [random system] seems friendlier, the inflection seems different."
115	"The [rule-based] system was unpredictable, while the [random system] was predictable."
117	"Both sounded alike but [the random system] sounder nicer and more natural."
200	"I do not perceive any differences between the systems."
202	"The different 'good job' after every right question seems more natural.".
204	"The [rule-based] system sounded more like a machine when saying 'good job'. The [random] system was more natural."
205	"The [rule-based] system sounded more natural but [the random system] hod more energy. I like the high-energy system."

Table 5.2: Responses by Subjects who Preferred the Rule-Based System

Subject	Reason for Preference
106	"Different tones of "good job" should be used [in the random system], to have a more natural feel."
107	"The responses were consistent, but [the rule-based system] seemed more familiar."
108	"I can't find any differences between the systems."
109	"Both systems seem equally helpful, there are no differences."
110	"I think the voice varied. The [random system] was robotic, but then [the rule-based system] sounded like it had some emotion."
111	"I felt that the responses were more varied with [the random system], but the [rule-based system] was more natural."
112	"There were minor glitches recognizing."
113	"Both systems sound very similar."
116	"They sounded almost the same, but not quite. At first I preferred the [random system] because I got almost all answers right, but after listening to the responses the [rule-based system] sounds more natural."
118	"The [rule-based system] sounded like it had more variety in the answers, while the [random system] sounded like it had planned the answers."
119	"I liked more the [rule-based system]. It sounded more like a person."

References

[1] K. S. D'Mello, S. Craig, A. Witherspoon, B. Mcdaniel, and A. Graesser. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adaptive Interaction*, 18(1–2):45–80, 2008.

[2] K. Forbes-Riley and D. Litman. Investigating human tutor responses to student uncertainty for adaptive system development. *ACII*, pages 678–689, 2007.

[3] D. Fossati. The role of positive feedback in intelligent tutoring systems. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Student Research Workshop*, Columbus, OH, 2008.

[4] A. Gravano and J. Hirschberg. Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, Brighton, UK, 2009.

[5] R. Higashinaka, K. Dohsaka, S. Amano, and H. Isozaki. Effects of quiz-style information presentation on user understanding. *Interspeech*, pages 2725–2728, 2007.

[6] T. Hollingsed. Responsive behavior in tutorial spoken dialogues. Master's thesis, The University of Texas at El Paso, 2006.

[7] T. Hollingsed and N. Ward. A combined method for discovering short-term affect-based response rules for spoken tutorial dialog. In *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTe)*, pages 61–64, 2007.

[8] J.F. Kelley. An empirical methodology for writing user-friendly natural language office information applications. *ACM Transactions on Office Information Systems*, 2(1):26–41, 1984.

[9] K. Porayska-Pomsta and H. Pain. Providing cognitive and affective scaffolding through teaching strategies: Applying linguistic politeness to the educational context. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, pages 77–86, 2004.

[10] K. Porayska-Pomsta and H. Pain. Diagnosing and acting on student affect: the tutor's perspective. *User Model. User-Adapt. Interact*, 18(1-2):125–173, 2008.

[11] T. Stocksmeier, S. Kopp, and D. Gibbon. Synthesis of prosodic attitudinal variants in german backchannel 'ja'. *Interspeech*, pages 1290–1293, 2007.

[12] N. Ward. *Didi, a Dialog Display and Analysis Tool.* World Wide Web, http://cs.utep.edu/nigel/didi/, 1998.

[13] N. Ward. Pragmatic functions of prosodic features in non-lexical utterances. *Speech Prosody*, 4:325–328, 2004.

[14] N. Ward and W. Tsukahara. A study in responsiveness in spoken dialog. *International Journal of Human-Computer Studies*, 59(6):959–981, 2003.

Appendix A

Spoken Dialogue Study
Questionnaire

Which system would you prefer to use?

A B

The purpose of this study was to investigate acknowledgement choice. How would you rate the overall naturalness of the acknowledgements produced by each system?

A: terrible **1** ….. 2 ….. 3 ….. **4** ….. 5 ….. 6 ….. **7** excellent

B: terrible **1** ….. 2 ….. 3 ….. **4** ….. 5 ….. 6 ….. **7** excellent

---------- answer the following questions after the experimenter plays back your dialogs ----------

How natural was each system?

A: terrible **1** ….. 2 ….. 3 ….. **4** ….. 5 ….. 6 ….. **7** excellent

B: terrible **1** ….. 2 ….. 3 ….. **4** ….. 5 ….. 6 ….. **7** excellent

How friendly was each system?

A: terrible **1** ….. 2 ….. 3 ….. **4** ….. 5 ….. 6 ….. **7** excellent

B: terrible **1** ….. 2 ….. 3 ….. **4** ….. 5 ….. 6 ….. **7** excellent

Which system would you prefer to use?

A B

Do you have any other comments on the systems?

Do you have any other comments about this experiment or anything else?

Appendix B

Spoken Dialogue Study
Agreement and Consent

Description of Study

This study is part of a research project, under the direction of Dr. Nigel Ward, aiming to improve the usefulness of spoken dialog systems. If you chose to participate, you will interact with a prototype of a tutoring system, then answer simple questions about the experience. This will take about 1 hour to complete. An audio record of the session will be made. Participants will receive class credit or be paid $10.00 for their participation. If at any point during the study a participant feels uncomfortable, they may contact Lola Norton, Institutional Coordinator for Research Review, at 747-5680. You may also contact Dr. Nigel Ward at 747-6827.

Participant Statement and Signature

I understand and agree that:

1. There are no known risks involved by participating in this study.

2. I may end my participation at any time and for any reason. I will be be given credit or payment for my participation whether or not I complete the experiment.

3. An audio recording will be made during the session.

4. My privacy will be protected: my name will not be associated with the experiment or made public in any way.

5. The results of the experiment, including audio records, may be used by members of the UTEP Interactive Systems Group and other persons designated by them for reasonable education, scientific, and technical purposes.

Signature: _____

Name: _____

Telephone/e-mail (if you are willing to be contacted if there is a follow-up study):

_____

Date: _____

I have received class credit or payment for my participation in this experiment: _____ (initial)

Experimenter Signature: _____

Appendix C

First Presidents of the United States

George Washington
He is ranked by scholars as one of the greatest U.S. Presidents

John Adams
He was the first vice president

Thomas Jefferson
Drafted the declaration of independence

James Madison
His wife's name was Dolly

James Monroe
Wrote a famous doctrine of the same name

John Quincy Adams
His father was also president

Andrew Jackson
Served as major general in the war of 1812

Martin Van Buren
Was nicknamed "the little magician"

William Henry Harrison
He was president for only 32 days

John Tyler
He had 15 children

Appendix D

List of Presidents in Reverse-Chronological Order

Jimmy Carter
He received the 2002 Nobel Peace Prize

Gerald Ford
His wife's name was Betty

Richard Nixon
He was the only president to resign from office

Lyndon Johnson
One of his most well-known programs was a Medicare amendment

John F. Kennedy
He was the 4th president to be assassinated while in office

Dwight Eisenhower
Established the interstate highway system

Harry Truman
His presidency saw the end of World War II

Franklin Roosevelt
Served the longest presidency in history

Herbert Hoover
He spoke Mandarin and Chinese

Calvin Coolidge
It is believed that he led the country into the Great Depression

Appendix E

Pref	Nat A	Nat B	Nat′ A	Nat′ B	Fri A	Fri B	Pref′	Rnd
B	5	6	5	7	6	7	B	A
B	6	7	6	7	7	7	B	B
B	6	7	7	7	7	7	B	A
B	5	6	6	6	5	6	B	B
B	6	6	6	7	6	7	B	A
A	7	4	7	7	7	7	B	B
A	6	7	6	7	7	7	B	A
B	6	6	5	6	6	6	A	B
A	6	6	6	7	6	7	B	A
B	6	7	7	7	7	6	A	B
B	3	5	5	4	3	5	B	A
A	6	5	6	4	7	5	A	B
B	3	4	4	5	4	5	B	A
A	7	6	7	7	7	7	A	B
A	6	6	6	5	7	6	A	A
B	4	6	2	6	5	5	B	B
B	2	5	4	6	6	7	B	A
B	3	6	3	5	2	6	B	B
B	5	6	4	6	5	6	B	A
A	6	4	6	5	7	7	A	B
A	7	7	6	7	7	7	A	A
A	6	2	6	2	7	2	A	B
B	5	6	5	6	6	6	A	A
B	6	7	5	4	5	5	B	B
A	5	4	5	3	6	3	A	A
A	5	4	6	7	5	6	B	B
B	6	7	4	6	5	7	B	A

Raw results. “Pref” represents pre-listening system preference. “Nat A” and “Nat B” represent naturalness ratings of systems A and B. “Nat′ A" and “Nat′ B” represent post-listening naturalness ratings of systems A and B. “Fri A” and “Fri B” represent post-listening friendliness ratings. “Pref′” represents post-listening system preference. “Rnd” indicates which was the random system.

Curriculum Vitae

Rafael Escalante was born in May 12, 1982 to Alma Ruiz Payán and Rafael Escalante Álvarez in Ciudad Juarez, Mexico. He graduated from Cathedral High School, El Paso, Texas in the summer of 2000 and entered the University of Texas at El Paso. In 2005 he joined the Interactive Systems Group as a Research Assistant while pursuing an undergraduate degree in Computer Science. He received his bachelor's degree in Computer Science in the summer of 2006.

It the summer of 2006, he entered the Graduate School of The University of Texas at El Paso. While pursuing a master's degree in Computer Science he worked as a Research and Teaching assistant.

Permanent address:

Atenea 5947, fracc. Minerva.

Cd. Juárez, Chihuahua, México. CP 32370.