

A Combined Method for Discovering Short-Term Affect-Based Response Rules for Spoken Tutorial Dialog

Tasha K. Hollingsed and Nigel G. Ward

Department of Computer Science
University of Texas at El Paso

tashah@gmail.com, nigelward@acm.org

Abstract

A good tutoring system should be able to detect and respond to subtle changes in the affective state of the learner, as a way to motivate and encourage the student, thereby improving the learning outcomes. This responsiveness should also operate at the sub-second timescale, as with some human tutors. Modeling this ability is, however, a challenge. This paper presents a combined method for the discovery of the rules governing such real-time responsiveness. This method uses both machine-learning and perceptual techniques, both with and without reference to internal states. This method is illustrated with the problem of choosing supportive acknowledgments in memory-reinforcing quiz dialogs. A wizard-of-oz experiment showed that users prefer a tutorial system based on responsive rules to one that chooses acknowledgments at random.

Index Terms: Responsiveness, Prosody, Emotion, User Modeling, Adaptation, Interpersonal Interaction, Word Choice

1. Introduction

A good conversationalist detects subtle, fleeting changes in their conversational partner's attitude and emotional state, and responds appropriately. For example, when you are speaking to someone who is excited over a job offer, you may respond with enthusiasm, increasing the pitch and energy in your speech, showing that you recognize their pleasure and are yourself happy for them. Similarly, speech systems should be able to detect changes in user attitude and respond appropriately, however interaction at this timescale is as yet poorly understood.

Good tutors in particular are often able to detect short-term learner states, such as uncertainty, pleasure, engagement, concentration and a desire for assistance. Responding appropriately can help improve the student's affective state and attitude towards the tutor, thereby increasing motivation [1].

For example, in the following dialog fragment,

Tutor: *Name ten of the fourteen countries in South America.*

Student: *Um, let's see . . . [long pause] uh, Argentina?*

Tutor: *Very good.*

Student: *And, um, Columbia?*

Tutor: *Good job, keep going.*

the tutor was able to detect the uncertainty in the student's speech and respond with encouragement to help the student persist.

In recent years there has been a growing understanding of how to model learner affect, attitude and cognitive state for more effective tutoring [3, 4, 5, 6, 7]. Since affect is expressed largely in the voice, achieving truly sensitive and supportive tutorial systems

may require systems that support spoken language interaction: indeed this is one of the primary motivations for building spoken tutorial systems. Work on this topic has examined what emotions are important, how they are expressed, and how tutors should respond [8, 9, 10, 11, 12].

This work builds most directly on [2], which showed that this sort of responsiveness can be of value in a tutorial context. That work showed how to use information in the student's speech — namely timing, pitch, and energy — together with contextual information to choose acknowledgments. The task was a simple memory game that asked participants to recall and name a sequence of Tokyo train stations. Users preferred the system which used inferred short-term affect, attitude and state to choose appropriate acknowledgments, over a version that produced acknowledgment at random.

In this work we had four aims. First, since this finding has not since been replicated, we wanted to do so. We felt this especially important since the finding was obtained in Japanese, a language and culture with a perhaps atypical interaction style. Second, more ambitiously, we wanted to show that this ability led not only to a user preference but also to improved learning. Third, we wanted to explore which dimensions of attitude and affect are most important for useful short-term responsiveness, to help delimit the phenomena that builders of future systems need to consider. Fourth, in light of the enormous amount of effort required to discover the rules of responsive interaction, we wanted to explore better development methods. This paper overviews the results; the details are elsewhere [13].

2. Domain and corpus

Following [2], we chose to develop our tutoring system using memory quiz tasks, as seen in the example above. These are representative of the type of review a student might do with a tutor if the aim is to memorize the times table, or the abbreviations of the chemical elements, or the codes to use for kinds of merchandise. We chose memory recall quizzes because they exhibit rich variation in acknowledgment use and a swift interplay between student and tutor, but are otherwise semantically and pragmatically quite simple. The quiz tasks included the countries of South America, the El Paso exits of Interstate 10, and the colleges of UTEP.

We then looked for an exemplary human tutor to model. We auditioned 12 students from the subject pool, in pairs. The one given the tutor role was supplied the answer lists, each a list of items that the student would need to say. The tutor was instructed to give hints as appropriate and the answers when necessary. The dialogs were recorded, and we chose as tutor the one who we felt was most friendly, flexible, and interesting to interact with.

We then collected the corpus, pairing our chosen tutor with 16 peers from the subject pool, doing 3 quiz tasks with each. During

⁰This research was sponsored in part by NSF Grant IIS-0415150. We thank David Novick, Will Enriquez, Guarav Garg and Ann Gates.

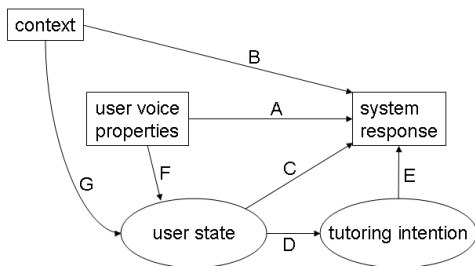


Figure 1: The Structure of the Inference and Response Choice Problem, rectangles indicating observables and ovals states.

the dialogues, the tutor was typically friendly and upbeat, maintaining a positive attitude throughout. If a student was having difficulty with a topic, she would give hints frequently and provide answers when needed.

As in the earlier study, we chose to focus on acknowledgment choice: the various ways in which the tutor responded to a correct guess. The corpus had 339 acknowledgments, of 41 types, with the most common being *very good* (accounting for 24% of the acknowledgments); *good job*; *uh-huh*; *mm-hmm*; *yeah*, *good job*; *yes*; *yeah*, *very good*; *yeah*; *right*; and *yay*, *good job*. To us this tutor’s acknowledgment use seemed rather atypical, both in the variety and the in the specific items used, but we accepted this as part of the interaction style which seemed to make her effective.

3. Rule discovery method

We set out to model the response behavior of the human tutor, in particular, the decision of which lexical form of acknowledgment to use in response to a correct answer, e.g. *very good* vs. *uh-huh*. In this study we did not examine the prosody of acknowledgments nor the question of when to give hints or the correct answer. Figure 1 shows the structure of the problem. We needed to discover what determined the tutor’s choice of acknowledgment. For purposes of building a system, the choice needed to be based on observables, namely the context and the user voice properties, however for purposes of analysis it was useful to consider two mediating variables: the user state and the tutor intention. The rest of this section describes the various methods used.

3.1. Subjectively judging how tutor responses relate to learner state

We first examined the data by repeated listening, trying to find commonalities among the contexts where each acknowledgment appeared. For this we came up with a few vague preliminary rules, based on correlations between context and properties of the user’s utterance, on the one hand, and tutor action on the other. We also began to identify some aspects of user state as probably relevant. This process corresponds to an initial exploration of relations A, B, and C in Figure 1.

3.2. Systematic correlation hunting

In the next stage we systematically explored some possible factors, of four types: 1. Dialog State Factors (7 factors), such as the number of incorrect guesses since the last correct answer, 2. Timing Factors (5), such as the time since the last tutor’s utterance, 3. Prosodic Features of the Student’s Answer (3), and 4. Other Factors (4). We labeled everything in the corpus so that these would

be easy to extract or derive. We then populated a spreadsheet with values for all 19 of these features for all 339 acknowledgments, and looked for combinations of these features able to predict the acknowledgment form. We did this by sorting the table on each of the independent variables in turn; this revealed four strong correlations, relating the acknowledgment chosen to: time into dialog, pitch slope on the correct answer, presence of *oh* before a correct answer, and to whether the student required a hint. This process gave a better understanding of relations A and B in Figure 1.

3.3. Initial formulation as a rule set

From these correlations we developed a first algorithm for acknowledgment choice, in the form of an ordered set of nine rules. This ruleset gave correct predictions 32% of the time. It’s not clear how bad this actually is, since it is not always the case that only one acknowledgment is appropriate in a given situation, however we felt that we could do better. In any case, this step gave an initial quantitative model for acknowledgment choice, based on relations A and B in Figure 1.

3.4. Decision tree learning

We next used decision tree learning, using XLMiner. This was a way to find an optional quantification of relations A and B in Figure 1. As test data we reserved one third of the data (one quiz, randomly chosen, from each subject), and used the rest for training. To avoid over-fitting we set the minimum number of items per bucket to 10. The decision tree obtained had only 6 decision nodes but a 41% match rate on the test data.

Some decisions used were easy to interpret. For example, the rule stating that if the pitch slope was $> .075$ then the acknowledgment to use is *very good* is easy to explain, as choosing to give an encouraging response when the learner is uncertain. Similarly, the rules stating that if a guess comes within .45 seconds of the previous guess it is to be acknowledged with *mm-hmm*, and if within 1.8 seconds of the previous correct guess with *uh-huh*, are easy to explain as choosing more perfunctory acknowledgments to the extent that the learner is doing better. However other decisions were hard to explain: in particular the tree chose between *good job* and *very good* based on fine distinctions in timing that had no obvious rationale.

3.5. Perceptual clustering of user states

We then took another pass at listening to the data, temporarily forgetting about acknowledgment choice and instead simply attempting to note the important characteristics of each learner guess, subjectively. We came up with a list of ten such characteristics. We then labeled each correct answer in the corpus with one or more of these. This process was not tightly structured; in terms of Figure 1, this aim was merely to elucidate the possible user states.

We then examined which of these labels tended to co-occur with specific acknowledgments. We discovered, for example, that when the learner produced a correct answer after some time doing poorly, the response was *very good* half the time. Some distinctions, however, remained elusive, for example, we were unable to completely explain the difference between *very good* and *good job*, and came to wonder whether our tutor was choosing between these two at random just for the sake of variety. In terms of Figure 1, this stage elucidated relation C.

3.6. Obtaining naive judgments

We then assembled some short dialog fragments, each about 60 seconds, by audio cut and paste, that either embodied these corre-

Rule	User Feeling/State /Response Characteristics	User Expression and Context	Tutor Intention	Tutor Expression
U1	Extremely Confident	No pause for thought	Allow user to continue without interrupting	none
U2	Confident	A correct guess followed by a short pause	Acknowledge, but don't interrupt	"uh-huh"
U3	Somewhat Confident	Short amount of time since user's last utterance and no hints needed	Acknowledge, but don't interrupt	"mm-hmm"
U4	Explicitly Uncertain	A correct guess followed by a long pause, current guess has rising pitch	Encourage user	"very good"
U5	Uncertain	Most recent guess followed by a short pause	Encourage user to continue and notify them that they did well	"good job"
U6	Uncertain	A very long time since last guess	Encourage user	"good job"
U7	First start	First guess	Start with an encouraging acknowledgment	"good job"
U8	Done	Final correct answer	Congratulate user on finishing the quiz	"good job"
U9	Not doing well	User needed many hints	Do not allow user to get discouraged	"very good"
U10	Doing better	User needed one hint	Notify user that they are doing well	"very good"

Figure 2: Unified Set of Rules for Acknowledgment Choice

lations or violated them, and had 8 naive subject pool members rate each acknowledgment on a scale of naturalness. Unfortunately this was not informative. Perhaps focus group analysis or other methods would have been better. However there was one incidental finding: the importance of prosody. This was seen in response to the question of what had influenced their ratings: half of the judges mentioned tone of voice, energy, and speed, even though we had explicitly requested them to rate whether the word itself was appropriate, not how it sounded. This suggests the examination of acknowledgment prosody as a priority for future work.

3.7. Rule unification

Although our various methods were certainly not converging on the same model, the resulting rulesets all seemed to be largely compatible. We therefore merged them, arriving at a unified set of rules. These related the observables (user action and context), the response (acknowledgment choice), and in addition the likely user state. In terms of Figure 1, this represented a unified account of relations A, B, and C.

Most of these rules were multiply motivated. Taken together, they incorporated most of the various insights uncovered in previous steps. Those left out related to rarer events, namely: learner excited or pleased at having hit on the right answer, learner got the right answer thanks to the hint, learner gave an unexpected but correct answer, learner self-corrected, and learner recycled an earlier guess (but now in the right place).

We then went through the rules and for each one added a rationale, in terms of what the tutor intention associated with that rule seemed to be. In terms of Figure 1, this added an account of relations D and E. The resulting composite set of rules is seen in Figure 2.

3.8. Rule quantification

Finally, we operationalized the rules. For example, we quantified Rule U1 (user explicitly certain) as applying whenever the previous guess was correct and the current guess occurred within 800

Quantitative Features of Context and User Expression
U7: First correct guess
U8: Last correct guess
U1: Preceded directly by another correct guess, intervening pause $\leq 800\text{ms}$
U4: Time since last correct guess $> 8\text{s}$ and pitch slope rises at least 40% over the last 250ms
U3: Time since last correct guess $\leq 5\text{s}$ and at least one filler and hints + incorrect guess ≤ 1
U2: Immediately after another correct guess, followed by a pause $\leq 5\text{s}$ and no hints, fillers, or incorrect guesses
U5: Time since last correct guess $\leq 8\text{s}$ and hints + incorrect guesses < 4
U6: Time since last correct guess $> 8\text{s}$ or the time since the last correct guess is $> 5\text{s}$ and incorrect guesses + hints ≥ 4
U10: Previously in state "Not Doing Well" (U9) and number of hints ≤ 1
U9: Incorrect guesses ≥ 2 or incorrect guesses + hints > 4

Figure 3: Revised and quantified conditions for each rule, rules ordered as checked

milliseconds of the previous guess. At this stage we also ordered the rules, in general putting the rules responding to clearer signals earlier. This process of quantification was done by hand, and was guided, somewhat arbitrarily, by the aim of maximizing the ability to predict the user-state labels. The results are seen in Figure 3.

In terms of Figure 1, this quantified relations F and G, and also, since each unified rule related all components of the model, the other relations. We then coded and tested this ruleset against the test data. Although the accuracy, 36%, was not much higher than on our first attempt, and lower than with the decision tree, we had more confidence in this ruleset; based on previous experience [2] we were less concerned with match to the corpus than with capturing important aspects of effective tutor behavior.

4. Evaluation method

In order to determine if the rules made a spoken-dialog tutoring system more usable and effective, we ran an experiment. We did this using Yesman [2], a Wizard-of-Oz (WOZ) spoken dialogue shell that in real time computes prosodic features in user utterances and detects the presence of fillers [14], while also keeping track of the dialogue state. Yesman contains no speech recognizer; it relies on an operator to determine whether each of the user's utterances is a correct answer or not. Yesman follows the embedded ruleset to govern hint giving and choice of acknowledgments.

Instructions, hints, and acknowledgments were in a recorded female voice. Acknowledgments were recorded with neutral intonation, being fairly flat in pitch and at unexceptional volumes and rates.

For the experiments we used as domains two sets of US presidents, the ten from Washington to Tyler and the ten from Coolidge to Carter. There were 28 subjects. For each set of presidents, they reviewed a study guide for three minutes, took a pre-quiz, used the tutorial system, and then took a post-quiz. The difference in scores between the two quizzes measured how much they had learned. After this they were given a brief transcript of their interactions with the systems, and were asked to listen to the recording while rating each acknowledgment on a scale from 1 (very natural) to 7 (very unnatural). Finally, after using and rating both systems, subjects filled out a questionnaire asking, among other things, which of the two systems they preferred overall.

Subjects used both the rule-based system and a random-acknowledgment control. We balanced the order of presentation and which system was used with which set of presidents. There were no significant interactions between the independent variables. All the subjects completed the dialogs and appeared to behave naturally in them.

5. Results

As hoped, subject ratings of the acknowledgments were higher with the rule-based system, averaging 2.41, vs. 2.71 with the random system (on a scale from 1, very human-like, to 7, not human-like at all), and the advantage, although small, was significant ($p < 0.05$ by a matched pairs t-test). In an overall comparison also, most subjects felt that the rule-based system was "more human-like" (11 of the 19 expressing a preference).

However, contrary to expectation, subjects' test scores improved more after using the random system than after using the rule-based system; the average improvement was .675 points (at one point per president) with the former and .57 with the latter, although the difference was non-significant (by a matched pairs t-test). Similarly, a bare majority felt that they would rather use the random-based system "to help you study for an important test" (11 of the 21 expressing a preference).

6. Discussion

The results confirm the finding of our earlier study: a system which uses affect-based response rules is perceived as more human-like. The preference was, however, somewhat weaker. Reasons may include weaker age-, language-, and cultural similarity between the speakers in the corpus and the subjects in the experiment, and use of subjects from a culture which relies more on explicit exchange of information and less on context and on subtle expressions of attitude and affect than does Japanese.

There was, however, no learning advantage. Other recent work has also failed to find a significant learning effect for this kind of modeling [12], raising the question of whether this sort of response

choice really has much value for this type of task.

The dimensions of user feeling and tutor intention identified as relevant here largely overlap those identified in other work, suggesting that there may be a small universal set of states, feelings and attitudes relevant to interaction at this timescale.

We believe that the discovery method employed here, using a variety of methods to gain insight on various views of the problem (Figure 1), and then combining them to come up with a unified set of rules, is likely to be of general value, especially when only a modest amount of data is available for analysis. The development of tools to support such analysis is a priority for future work.

7. References

- [1] Lepper, M. R. and R. W. Chabaty. 1988. Socializing the intelligent tutor: Bringing empathy to computer tutors. In H. Mandl and A. Lesgold (eds.) *Learning issues for intelligent tutoring systems*. Springer Verlag.
- [2] Ward, Nigel and Wataru Tsukahara. 2003. A Study in Responsiveness in Spoken Dialog. *International Journal of Human-Computer Studies*, 59 (6), 959-981.
- [3] Heylen, Dirk and Anton Nijholt and Rieks op den Akker. 2005. Affect in Tutoring Dialogues. *Applied Artificial Intelligence*, 19, 287-311.
- [4] Beck, Joseph E. 2005. Engagement Tracing: Using response times to model student disengagement. *Proceedings of Artificial Intelligence in Education*, 88-95.
- [5] Alexander, Samuel and Stephen Hill and Abdolhossein Sarrafzadeh. 2005. How do Human Tutors Adapt to Affective State? in *Proceedings of the User Modeling '05 Workshop on Adapting the Interaction Style to Affective Factors*.
- [6] Aist, Gregory, Barry Kort et al. 2002. Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor That Listens Increases Student Persistence. *Proc. 6th Intl. Conf. on Intelligent Tutoring Systems*, pg 992.
- [7] Elliott, Clark, Jeff Rickel and James Lester. 1999. Life-like Pedagogical Agents and Affective Computing: An Exploratory Synthesis. in *Artificial Intelligence Today*, eds M. Wooldridge and M. Veloso, Springer, pp 195-212.
- [8] Graesser, A. C., K. Wiemer-Hastings, P. Wiemer-Hastings, and R. Kreuz. 1999. Auto Tutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, 1, 35-51.
- [9] Liscombe, Jackson, Julia Hirschberg, and Jennifer J. Venditti. 2005. Detecting Certainty in Spoken Tutorial Dialogues. In *Proc. Interspeech*.
- [10] Litman, Diane and Kate Forbes-Riley. 2004. Predicting Student Emotions in Computer-Human Tutoring Dialogues. *Proc. 42nd ACL*.
- [11] Forbes-Riley, Kate, Diane Litman et al. 2007. Exploring Affect-Context Dependencies for Adaptive System Development. *Proc. NAACL-HLT 2007*.
- [12] Pon-Barry, Heather, Karl Schultz et al. 2006. Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems *International Journal of Artificial Intelligence in Education*, 16, 171-194.
- [13] Hollingsed, Tasha K. 2006. Responsive Behavior in Tutorial Spoken Dialogues. University of Texas at El Paso, Department of Computer Science, Masters Thesis.
- [14] Garg, Gaurav and Nigel G. Ward. 2006. Detecting Filled Pauses in Tutorial Dialogs. *Tech. Report UTEP-CS-06-32*.