

# Interactions between Speech Recognition Problems and User Emotions

Mihai Rotaru, Diane J. Litman and Katherine Forbes-Riley

Department of Computer Science  
University of Pittsburgh, 210 S. Bouquet, Pittsburgh, PA, 15260, USA  
{mrotaru, litman, forbesk}@cs.pitt.edu

## Abstract

Understanding how speech recognition problems affect the interaction with the user is a topic of great interest for the spoken dialogue community. In this paper, we examine the dependencies between speech recognition problems in adjacent turns. We also examine the dependencies between speech recognition problems and student emotions within a turn and in adjacent turns. We apply Chi Square ( $\chi^2$ ) analysis to a corpus of speech-based computer tutoring dialogues to discover these dependencies. We find that rejections are followed by more rejections than expected if there was no dependency between rejections, and that misrecognitions are followed by more misrecognitions than expected. We also find a strong dependency between recognition problems in the previous turn and user emotion in the current turn: after a system rejection there are more emotional user turns than expected. Surprisingly, in our data, we find no relationship between user emotions and recognition problems within a turn nor between previous turn user emotions and current turn recognition problems.

## 1. Introduction

Previous work has highlighted the impact of speech recognition problems on various dialogue phenomena. In reaction to system misrecognitions, users try to correct the system by employing strategies that work in human-human interactions. They tend to correct the system by switching to a prosodically marked speaking style [1] in many cases consistent with hyperarticulated speech [2]. Since most recognizers are not trained on this type of speech [3], these attempts lead to further errors in communication [1, 2]. The resulting “chaining effect” of recognition problems can affect the user emotional state; a frustrated and irritated user will lead to further recognition problems [4]. Ultimately, the number of recognition problems is negatively correlated with the overall user satisfaction [5].

These findings suggest dependencies between recognition problems, and also dependencies between recognition problems and user emotions. We use  $\chi^2$  analysis to discover these dependencies in a corpus of speech-based computer tutoring dialogues.

Our work extends previous research in several aspects. We extend the interaction studies by looking at several types of recognition problems capturing aspects like automatic speech recognition (ASR) performance and whether the recognition problem was perceived by the user. We do not restrict our analysis to corrections [2] or to corrections through repetition [1] but study the dependencies on all turns from our corpus. We look at speech-based computer tutoring dialogues instead of more commonly used information retrieval dialogues [1, 2]. In information retrieval tasks, users

are in general familiar with the task and can easily recognize when the system made a mistake (e.g. the system misrecognized the departure city in the air travel domain). In contrast, in tutoring dialogues users are not very familiar with the domain and might not detect a recognition problem immediately. For example, in many cases users do not know the right answer to a computer tutor question. Finally, we extensively study the interaction between recognition problems and user emotions. Because of the tutoring domain, our data also contains emotion types that are absent or very rare in information retrieval dialogues (e.g. uncertain, enthusiastic, etc).

First we look at the effect of recognition problems in the previous turn on the recognition problems in the current turn. As suggested by previous work [1, 2], we expect a “chaining effect”: recognition problems trigger more recognition problems. Our analysis yields a chaining effect for ASR misrecognitions and system rejections.

Next, we study interactions between recognition problems and user emotions. Previous work [6, 7] has shown that user emotions expressed through the speech channel can be coded and predicted reliably. Detecting user emotions allows a system to adapt its strategies in accordance to user state. Because certain user emotions should be avoided (e.g. angry, frustrated), it is also important to understand what causes users to exhibit those emotions. We hypothesized that a recognition problem in the previous turn will affect the user emotion state in the current turn and, indeed, we find that after rejections users are more emotional than expected. Surprisingly, there are less emotion turns after a semantic misrecognition. We anticipated an interaction between user emotion and recognition problems within a turn but, in contrast with [4], our data does not support this hypothesis. We also find no dependencies between user emotions in the previous turn and recognition problems in the current turn.

## 2. Corpus and annotation

The corpus analyzed in this paper consists of 100 experimentally obtained spoken tutoring dialogues between 20 students and **ITSPOKE** (Intelligent Tutoring **SPOKE**n dialogue system). ITSPOKE [8] is a speech-enabled version of the text-based Why2-Atlas conceptual physics tutoring system [9]. When interacting with ITSPOKE, students first type an essay answering a qualitative physics problem using a graphical user interface. ITSPOKE then engages the student in *spoken* dialogue (using head-mounted microphone input and speech output) to correct misconceptions and elicit more complete explanations, after which the student revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision. Because speech recognition is imperfect, after the data was collected, each student utterance

in our corpus was manually transcribed by a project staff member. An annotated excerpt from our corpus is shown in Figure 1 (punctuation added for clarity). The excerpts show both what the student said (the STUDENT labels) and what ITSPOKE recognized (the ASR labels). The excerpt is also annotated with concepts that will be described next.

## 2.1. Speech recognition problems

One form of recognition problem is the **Rejection**. Rejections occur when ITSPOKE is not confident enough in the recognition hypothesis and asks the student to repeat (Figure 1, STUDENT<sub>1</sub>). For our  $\chi^2$  analysis, we define the **REJ** variable with two values: **Rej** (a rejection occurred in the turn) and **noRej** (no rejection occurred in the turn). Not surprisingly, ITSPOKE also misrecognized some student turns. When ITSPOKE heard something different than what the student actually said *but was confident* in its hypothesis, we call this an **ASR Misrecognition** (a binary version of the commonly used Word Error Rate) (Figure 1, STUDENT<sub>2,4</sub>). Similarly, we define the **ASR MIS** variable with two values: **AsrMis** and **noAsrMis**.

---

**ITSPOKE:** That's great. So, once again on the original question. Let's think about the forces involved if you hit a brick wall with your fist. There is a force exerted by your fist on the wall at the time it hits the wall. In what direction is this force relative to the wall?

**STUDENT<sub>1</sub>:** Towards it (*Neutral*)

**ASR<sub>1</sub>:** words newton (*Rej*)

**ITSPOKE:** Could you please repeat that?

**STUDENT<sub>2</sub>:** towards it (*Emotional; Tut*) [*correct*]

**ASR<sub>2</sub>:** towards it moves (*AsrMis*) [*correct*]

**ITSPOKE:** Yeah. What happens to your fist when it hits the wall?

**STUDENT<sub>3</sub>:** it hurts (*Neutral*) [*correct*]

**ASR<sub>3</sub>:** it hurts [*correct*]

**ITSPOKE:** Good. What force are you feeling when your fist hits the wall?

**STUDENT<sub>4</sub>:** the wall pushing back on you (*Neutral*) [*correct*]

**ASR<sub>4</sub>:** ball pushing button new (*AsrMis, SemMis*) [*incorrect*]

---

Figure 1: Human-Computer Dialogue Excerpt.

Semantic accuracy is more relevant for dialogue evaluation, as it does not penalize for word errors that are unimportant to overall utterance interpretation. In the case of form-based information access spoken dialogue systems, computing semantic accuracy (also known as Concept Accuracy) is straightforward: its value is the percentage of the concepts conveyed by the user in the turn that were correctly recognized. In contrast, in the tutoring domain there are no clear forms with slots to be filled. We base our semantic accuracy on the “correctness” measure of the student turn. For each student turn, ITSPOKE interprets it and labels its correctness with regard to whether the student correctly answered the tutor question (see the labels between square brackets in Figure 1). We define **Semantic Misrecognition** as cases where ITSPOKE was *confident* in its recognition hypothesis and the correctness interpretation of the

recognition hypothesis is different from the correctness interpretation of the manual transcript (Figure 1, STUDENT<sub>4</sub>). Similarly we define the **SEM MIS** variable with two values: **SemMis** and **noSemMis**. The top part of Table 1 lists the distribution for our three recognition problem variables on the entire corpus as well as on the emotion tagged subset discussed in the next subsection.

Table 1: Variable distributions in the human-computer corpus and the emotion tagged subset.

Variable	Value	Full corpus (2386)	Emo. subset (337)
<b>Speech recognition problems</b>			
ASR	AsrMis	25.1%	24.9%
MIS	noAsrMis	74.9%	75.1%
SEM	SemMis	5.7%	6.2%
MIS	noSemMis	94.3%	93.8%
REJ	Rej	6.8%	5.3%
	noRej	93.2%	94.7%
<b>Student emotion</b>			
EnE	Emotional	N/A	55.5%
	Neutral	N/A	44.5%
<b>Emotion source</b>			
EMO	Phy	N/A	30.6%
	Tut	N/A	12.8%
	Both	N/A	12.2%
	Neutral	N/A	44.5%

## 2.2. Emotion annotation and emotion source

We have developed an annotation scheme for annotating student emotions and attitudes during tutoring [6]. This scheme was previously applied to a *subset* of the human-computer corpus discussed above (15 dialogues from 10 students resulting in 337 student turns). In our annotation scheme, each student turn is labeled for both strong and weak perceived expressions of emotion (confused, bored, irritated, uncertain, confident, enthusiastic, etc.). For our experiments we will use the consensus labeling of this data (the two original annotators revisited each originally disagreed case and, through discussion, sought a consensus label).

In our previous work, we studied various combinations of our emotion classes [6]. Here we use the emotional/non-emotional (EnE) distinction. Turns where the student exhibited any emotion are labeled as emotional (Figure 1, STUDENT<sub>2</sub>). All other turns are labeled as non-emotional (Figure 1, STUDENT<sub>1,3,4</sub>). For our  $\chi^2$  analysis, we define the **EnE** variable (see Table 1) with two values: **Emotional** and **Neutral** (non-emotional).

Similarly to the task and communicative information level used in the DAMSL dialogue act tagging scheme [10], we also tagged each student turn as to whether the emotions expressed in that turn were responding to the tutored domain, to the interaction with the computer tutor or to both. For our analysis, we define a new variable **EMO SRC**: the emotion source (Table 1). All emotional turns where the emotion was responding to the tutored domain of physics were labeled as **Phy**; emotions like uncertain, confused or confident are usually expressed in these turns. Turns where the emotion was responding to the interaction with the computer tutor were labeled as **Tut** (Figure 1, STUDENT<sub>2</sub>); in general the student was irritated or bored in these turns. All other

emotional turns were labeled as **Both** since both sources for emotions were present. Non-emotional turns have no emotion source and were labeled as **Neutral**.

### 3. $\chi^2$ analysis

To discover the dependencies between our variables, we apply the  $\chi^2$  test.  $\chi^2$  is a non-parametric test of the statistical significance of the relationship between two variables. We illustrate our analysis method on the interaction between rejections in the previous turn (REJ(-1)) and rejections in the current turn (REJ). In our notations, whenever a variable name or value is followed by a “(-1)” it refers to the previous turn variable/value; otherwise it refers to the current turn.

The  $\chi^2$  value assesses whether the differences between observed and expected counts (in parentheses) are large enough to conclude a statistically significant dependency between the two variables (Table 2). The observed counts are computed from the data. For example, there are 91 cases of a non-rejection after a rejection (Figure 1, STUDENT<sub>2</sub> turn is one of them). The expected counts are the counts that would be expected if there were no relationship at all between the two variables.  $\chi^2$  value would be 0 if observed and expected counts were equal. To account for a given table’s degree of freedom and one’s chosen probability of exceeding any sampling error, the  $\chi^2$  value has to be larger than the critical  $\chi^2$  value. For Table 2, which has one degree of freedom, the critical  $\chi^2$  value at a  $p < 0.05$  is 3.84. Our  $\chi^2$  value of 363.49 greatly exceeds this critical value. We thus conclude that there is a statistically significant dependency between rejections in the previous turn and in the current turn.

Table 2: REJ(-1) – REJ interaction.

REJ(-1) \ REJ	noRej	Rej
noRej(-1)	2036 (1976.3)	89 (148.7)
Rej(-1)	91 (150.7)	71 (11.3)
$\chi^2$ value: 363.49		$p < 0.0001$

A comparison of observed counts and expected counts in a significant dependency can give insight on how the interaction works. For example, in Table 2, after a rejection, we see *more* rejections than expected (71 instead of 11.3) and less non-rejections (91 instead of 150.7).

## 4. Results

### 4.1. Interactions between recognition problems

We discover three significant dependencies using the data from the full corpus. We find a strong dependency between a *rejection* in the previous turn and a rejection in the current turn (as discussed in Section 3 and reported in Table 2). Our data indicates a chaining effect for rejections: there are more rejections after a rejection than expected.

Our other two dependencies involve an *ASR misrecognition* in the previous turn (Table 3). After those turns we find more turns with ASR misrecognition than expected, suggesting again a chaining effect in ASR recognition problems. Surprisingly, we find *fewer* rejections after an ASR misrecognition than expected.

Since the analysis in Section 4.2 will be restricted to the emotion tagged subset, we wanted to understand whether this subset is a good sample of the full corpus. We already saw in Table 1 that the recognition problem distributions are similar.

We also applied the  $\chi^2$  analysis to the emotion subset of our corpus to see whether the results from this section can be discovered by the emotion subset. In this subset, the REJ(-1) – REJ dependency discovered in the full corpus is also significant ( $p < 0.0001$ ); the ASR MIS (-1) – ASR MIS dependency from the full corpus is almost a trend in this subset ( $p < 0.1131$ ) while the ASR MIS (-1) – REJ dependency is not that strong ( $p < 0.1821$ ). But most importantly, we do not find any new dependencies; all trends or significant dependencies discovered in this subset are also trends or significant in the full corpus.

Table 3: ASR MIS(-1) – ASR MIS and ASR MIS(-1) – REJ interactions.

ASR MIS(-1) \ ASR MIS	noAsrMis	AsrMis
noAsrMis(-1)	1305 (1274.8)	394 (424.2)
AsrMis(-1)	411 (441.2)	177 (146.8)
$\chi^2$ value: 11.14		$p < 0.0008$

  

ASR MIS(-1) \ REJ	noRej	Rej
noAsrMis(-1)	1567 (1580.1)	132 (118.9)
AsrMis(-1)	560 (546.9)	28 (41.2)
$\chi^2$ value: 6.07		$p < 0.0137$

### 4.2. Emotions - recognition problems interactions

The analysis in this section is restricted to the emotion tagged subset. We begin by looking at the interaction between system recognition errors in the previous turn and the user emotion in the current turn. We find a strong dependency ( $p < 0.0007$ ) between a *previous rejection* and the emotion of the current turn. As expected, our results indicate that after a rejection there are more emotional turns than expected (see Table 4). We were also interested to understand the valence and the source of the larger number of emotional turns after a rejection. Unsurprisingly, we found that the majority of these emotions correspond to negative emotions (14 out of 17).

Table 4: REJ(-1) – EnE, REJ(-1) – EMO SRC and SEM MIS(-1) – EnE interactions.

Rej(-1) \ EnE	Emotional	Neutral
noRej(-1)	163 (169.9)	141 (134.1)
Rej(-1)	17 (10.1)	1 (7.9)
$\chi^2$ value: 11.49		$p < 0.0007$

  

REJ(-1) \ EMO SRC	Tut+Both	Phy	Neutral
noRej(-1)	67 (78.3)	96 (91.6)	141 (134.1)
Rej(-1)	16 (4.7)	1 (5.4)	1 (7.9)
$\chi^2$ value: 39.71		$p < 0.0001$	

  

SEM MIS(-1) \ EnE	Emotional	Neutral
noSemMis(-1)	172 (168.3)	129 (132.7)
SemMis(-1)	8 (11.7)	13 (9.3)
$\chi^2$ value: 2.89		$p < 0.0891$

Moreover, we find a strong dependency between a rejection and the source of the emotion in the next turn. This dependency manifests in the fact that the source of the emotion is due to the interaction with the tutor (*Both+Tut*)<sup>1</sup>

<sup>1</sup> We had to combine the *Tut* and *Both* values because the small number of expected count for each value after a rejection was violating the  $\chi^2$  test requirements.

more than expected and less due to the physics domain (*Phy*) (see Table 4). These findings stress the negative impact of system rejections on student emotions. Together with the chaining effect observed in Section 4.1, it motivates future studies on efficient strategies for handling rejections in a spoken dialogue system.

There is also a trend for a dependency between *semantic misrecognition* in the previous turn and emotions in the current turn ( $p < 0.0891$ ). Surprisingly, after a semantic misrecognition there are *fewer* emotional turns than expected (see Table 4).

Next, we look at the interaction between emotions and recognition problems *within a turn*. It is widely believed that emotional user turns will be harder to recognize. This belief stems mainly from the fact that many acoustic models are not trained on emotional speech [3] (e.g. frustrated, angry, etc). Surprisingly, we find no significant dependencies between emotions and any of the recognition problem types.

Finally, we look at how user emotional state in the *previous turn* can affect the recognition of the current turn. For example, a student that became frustrated in the previous turn might develop a negative reaction to the tutor prompts. As a result of that, the student might change her speaking style and potentially influence the recognition performance. Again, we discover no dependencies between the emotion of the previous turn and the computer tutor ability to recognize current student turn.

## 5. Discussion and future work

In this paper we investigate dependencies between speech recognition problems, and dependencies between speech recognition problems and user emotions. We find that after rejections there are more rejections than expected and more emotional turns than expected. The emotions from these emotional turns originate from the interaction with the computer tutor more than expected and from the tutored domain less than expected. We also find that after an ASR misrecognition there are more ASR misrecognitions than expected but fewer rejections. Surprisingly we find that our user emotions classes have no effect on recognition problems.

We believe that an important factor in interpreting our findings is the type of tutoring dialogues in our corpus. Previous work has studied interactions between recognition problems in information retrieval dialogues [1, 2]. In these dialogues, the user usually needs to correct a SemMis. Thus both SemMis and Rej affect the “normal” dialogue flow, resulting in the chaining effect for the combination of rejections and semantic misunderstandings reported in [1, 2]. In contrast, in our tutoring dialogues, the dialogue can continue “normally” even after a SemMis: if the user’s answer was correct but misrecognized as incorrect, ITSPOKE will launch an additional subdialogue to correct the student “mistake” or give away the correct answer. If the user’s answer was incorrect but misrecognized as correct, ITSPOKE will just move to the next issue in the tutoring agenda. We hypothesize that this difference explains why we find a chaining effect for REJ but no effect for SEM MIS.

We also hypothesize that ITSPOKE’s behavior after interpreting a correct user answer as incorrect delays or inhibits user emotional response to the SemMis; this could explain why after a SemMis we find less emotional turns than expected. Coupled with the negative effect of rejections, it

suggests that for our tutoring task, a *lower* rejections threshold might be more beneficial (at least with respect to reducing the number of emotional student turns). This is on par with the results reported in [11]. They show that the most effective *human* strategy when faced with recognition problems is not to signal misunderstanding but to ask task related questions. By engaging in additional tutoring subdialogues after the additional SemMis produced by a lower rejection threshold, our tutor will exhibit a similar behavior.

Extending [1, 2], we also look at the interactions with ASR MIS and discover a chaining effect for ASR MIS. We also find fewer rejections after an ASRMis. We are currently examining our data for interpretations of these dependencies.

[4] show that frustration interacts with recognition problems. In contrast, we do not find any interactions between emotions and recognition problems. However, we only looked at the emotion/non-emotional distinction. In our future work, we plan to experiment with each emotion label separately (e.g. uncertain, irritated, frustrated) and with other coarse-grain emotion labels (e.g. positive, negative; see [6])

In our data we have also annotated how the emotion was conveyed: prosodically, lexically or both. We hope to gain more insight on the role of user emotions by incorporating this information in our future analyses. Ultimately, we expect that these analyses and their findings will help us design more adaptive spoken dialogues systems.

## 6. Acknowledgements

This research is supported by NSF Grant No. 0328431.

## 7. References

- [1] G. Levow, "Characterizing and recognizing spoken corrections in human-computer dialogue," COLING-ACL, 1998.
- [2] M. Swerts, D. Litman, and J. Hirschberg, "Corrections in Spoken Dialogue Systems," ICSLP, 2000.
- [3] H. Soltau and A. Waibel, "Specialized acoustic models for hyperarticulated speech," ICASSP, 2000.
- [4] A. Boozer, S. Seneff, and M. Spina, "Towards Recognition of Emotional Speech in Human-Computer Dialogues," CSAIL Research Abstract 2003.
- [5] M. Walker, R. Passonneau, and J. Boland, "Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems," ACL, 2001.
- [6] D. Litman and K. Forbes-Riley, "Annotating Student Emotional States in Spoken Tutoring Dialogues," SIGdial Workshop on Discourse and Dialogue (SIGdial), 2004.
- [7] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to Find Trouble in Communication," *Speech Communication*, vol. 40 (1-2), 2003.
- [8] D. Litman and S. Silliman, "ITSPOKE: An intelligent tutoring spoken dialogue system," HLT/NAACL, 2004.
- [9] K. VanLehn, P. W. Jordan, C. P. Rosé, D. Bhembé, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, and R. Srivastava, "The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing," *Intelligent Tutoring Systems (ITS)*, 2002.
- [10] M. Core and J. Allen, "Coding Dialogs with the DAMSL Annotation Scheme," AAAI Fall Symposium on Communicative Action in Humans and Machines, 1997.
- [11] G. Skantze, "Exploring Human Error Handling Strategies: Implications for Spoken Dialogue Systems," ISCA Workshop on Error Handling in Spoken Dialogue Systems, 2003.