Grounding Emotions in Human-Machine Conversational Systems

Giuseppe Riccardi and Dilek Hakkani-Tür

AT&T Labs–Research, 180 Park Avenue,
Florham Park, New Jersey, USA
{dsp3, dtur}@research.att.com

Abstract. In this paper we investigate the role of user emotions in human-machine goal-oriented conversations. There has been a growing interest in predicting emotions from acted and non-acted spontaneous speech. Much of the research work has gone in determining what are the *correct* labels and *improving* emotion prediction accuracy. In this paper we evaluate the *value* of user emotional state towards a computational model of emotion processing. We consider a binary representation of emotions (positive vs. negative) in the context of a goal-driven conversational system. For each human-machine interaction we acquire the temporal emotion sequence going from the initial to the final conversational state. These traces are used as features to characterize the user state dynamics. We ground the emotion traces by associating its patterns to dialog strategies and their effectiveness. In order to quantify the *value* of emotion indicators, we evaluate their predictions in terms of speech recognition and spoken language understanding errors as well as task success or failure. We report results on the $11.5K$ dialog corpus samples from the *How may I Help You?* corpus.

1 Introduction

In the past few years, there has been a growing interest in the speech and language research community in understanding the paralinguistic channel in human-machine communication. The paralinguistic component includes such information as speaker's age, gender, speaking rate and state. In this paper we are going to address the latter, that is the emotional state of users engaged in goal oriented human-machine dialogs.

In goal-oriented human machine communication the user might display different states due to prior conditions (e.g. previous attempts at solving the task) or poor machine cooperativeness in acknowledging and/or solving a problem (e.g. machine misunderstanding) or poor reward (e.g. the dialog is not successful). Prior conditions affect the user state in a way that can be detrimental or beneficial to the outcome of the interaction. Being able to detect users' emotional state is crucial and would require to be able to know or estimate the *profile* of the user. Cooperativeness in a human-machine dialog allows the machine to elicit important task-related information at the early stages of the interaction

M. Maybury et al. (Eds.): INTETAIN 2005, LNAI 3814, pp. 142–152, 2005.

and/or resolve problematic turns due to speech recognition, understanding and language generation performance. The outcome of the interaction will impact the final user state which will probably *re-emerge* later in time.

In this paper we analyze the role of emotions in human-machine dialogs by grounding them into the action-reaction traces of human-machine dynamics. Previous work emphasis has been on detecting and predicting emotion (positive/negative) states [1,2,3]. This paper is aimed at analyzing the effect of emotional patterns and its impact on machine reactions and performance. The fundamental issues we are going to address are:

- The impact of the user state on machine's dialog strategies.
- The impact of the user state on machine's performance.
- The impact of the user prior conditions on future state transitions.

The measures used to quantify the *impact* are done along two dimensions. The first is machine accuracy in recognizing, understanding and managing user's spoken input. We show that user state has serious effect on the accuracy of state-of-the-art models which are based on statistical models (speech recognition and understanding) or hand-crafted machine action strategies (dialog manager). The second dimension is based on the success or failure of the user-machine dialog in accomplishing the user's task. While the granularity of the first dimension is at the utterance or sub-dialog level the second dimension is a measure of the complete sequence of user-machine exchanges.

As most of thestate-of-the-art spoken dialog systems, the current system is emotionless. Thus the analysis that will be carried out in this paper will aim at *exposing* the emotional component of the user state in such class of conversational machines. The ultimate goal is to provide a set of parameters or machine actions that could benefit from using emotion indicators.

We provide a statistical analysis based on the *How May I Help You?* spoken dialog database [4]. The database includes transcriptions of spoken utterances, transcriptions of system prompts, semantic tags (user intent or calltype) estimated by the machine and labeled by a human, dialog acts and manually labeled emotion tags. In the following section we describe the database, its annotation labels and protocol. In Sections 3, 4, 5 we quantify the relations between emotion patterns, machine behavior and performance respectively.

2 The *How May I Help You?* Spoken Dialog System

"How May I Help You?^{SM}", AT&T's natural language human-computer spoken dialog system, enables callers to interact verbally with an automated agent. Users may ask for their account balance, help with calling rates and plans, explanations of bill charges, or identification of numbers on bills that they do not recognize. The machine is expected to understand their requests and route them to the correct information. If the system wants to confirm or clarify a customer's response, the dialog manager asks for more information; if it is still not clear, it routes the caller to a service representative. Speech data from the deployed

System:	How may I help you?
User:	I need to find out about a number that I don't recognize.
System:	Would you like to look up a number you don't recognize on your bill?
User:	Yes I would.
System:	Are you calling from your home phone?
User:	Yes I am.
System:	...

Fig. 1. Sample dialog from the HMIHY Corpus

"How May I Help You?^{SM}" system has been compiled into a corpus referred to as HMIHY [4,5]. Figure 1 presents the transcription of an example dialog from the corpus.

In the HMIHY spoken dialog system the machine is trained to perform large vocabulary Automatic Speech Recognition (ASR) based on state-of-the-art statistical models [6]. The input spoken utterance is modeled as a sequence of acoustic and lexical hidden events (acoustic and language models). The word sequence output from the ASR module is then parsed to determine the user intent (calltype) using robust parsing algorithms [4]. This spoken language understanding (SLU) step provides a posterior probability distribution over the set of intents. While the speech recognition and robust parsing models are trained off-line, the posterior distribution is computed on-line from the spoken input. The posterior probabilities are used by the Dialog Manager (DM) to infer the most appropriate system dialog act. The algorithm used by the DM is heuristic-based and partially domain-dependent. The DM algorithm principle is designed to cope with ASR and SLU errors and converge to a dialog final state [7].

2.1 System Performance Metrics

The ASR performance is evaluated using the standard word error rate (WER) measure. The SLU performance is evaluated using top class error rate (TCER), which is the percentage of utterances where the top-scoring calltype output of SLU is not among the true call-types labeled by a human. In order to evaluate the dialog level performance, three labelers labeled a 747 dialog subset of the HMIHY corpus with three labels: *Task Failure*, *Task Success*, and *Other*. The labelers were given the instructions to label each dialog with one of these labels, using the prompt transcriptions, user response transcriptions, system call-types, and human labeler call-types for each utterance. We used the first 100 dialogs in order to compare the errors and strategies among the three labelers, and converge to stable annotation guidelines. In the comparisons in this paper, we only use the following 647 dialogs. For the initial 100 dialogs, Cohen's Kappa statistic was 0.42 for the three labelers, and for the final 647 dialogs it was 0.52, showing an improvement in the labeling guidelines.

2.2 Corpus Description and Annotation

We have annotated the HMIHY corpus in two phases. In the first phase [8], 5,147 user turns were sampled from 1,854 HMIHY spoken dialogs and annotated with one of seven emotional states: *positive/neutral, somewhat frustrated, very frustrated, somewhat angry, very angry, somewhat other negative, very other negative*. Cohen's Kappa statistic, measuring inter-labeler agreement, was calculated using this data set. A score of 0.32 was reported using the full emotion label set whereas a score of 0.42 was observed when the classes were collapsed to *positive/neutral* versus *other*. The small emotion label set is not equivalent to the larger one, but it provides us with more consistently labeled data.

In the first phase we have encountered a high degree of variability (with respect to the number of labels) and unreliability (annotator agreement). In the second phase we have quantized the emotion labels into two labels, tokenized the corpus in terms of complete dialogs and increased the size of the corpus to 11,506 complete dialogs (40,551 user turns). Each new user turn was labeled with one of the emotion labels mentioned above. We used this expanded corpus labeled with *positive* versus *negative* user states for the experiments presented in this paper. In 8,691 dialogs, all turns are labeled as *positive*, and 2,815 dialogs have at least one turn labeled as *negative*. 35,734 of the user turns are labeled as *positive*, and the rest (4,817) of the user turns are labeled as *negative*.

3 Emotions and Machine Behavior

In modeling the user state $s(t)$ we assume that there is a component dependent on *prior conditions* and a component dependent on the dynamic performance of the human-machine interaction. In the next sections, we investigate the relations between each component and user transcriptions, semantic and dialog annotations. For each component we evaluate the effect on system behavior and performance. The number of positive (negative) state labels in a complete dialog trace will be indicated with p (n). The value of a state label $s(t)$ at time t (i.e., turn $t+1$) is 0 (1) for positive (negative) labels.

3.1 Empirical Distributions over Time

As most of the current state-of-the-art spoken dialog systems, the HMIHY system is *emotionless* both from the input (detection) and output (generation) side. The DM representation of the user state is defined only in terms of dialog act or expected user intent. On the hand, in the following analysis we investigate how the observed emotional component of the user state impacts such a system.

The first question that we address is on the state probability over time to be in a negative state. During the interaction the machine processes noisy input (e.g., speech recognition errors) and makes an estimate of the noise level. There are two types of DM strategies that would exploit this noise estimate at each turn (intent posterior probability). The first is to assume that the information

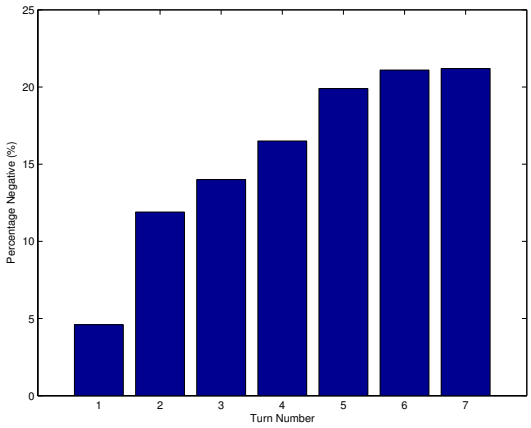

Fig. 2. Percentage of negative turns over time (turn number) within a dialog. The histogram is truncated at $t = 7$.

acquired is correct and act accordingly. The second is to assume the input is not correct (partially or totally) and apply an error recovery DM strategy.

Fig 2 shows the percentage of spoken utterances with negative emotions over time (dialog turns). The monotonic increase of $P(s(t) = 1)$ over the course of the dialog can be explained in two different ways. First, at each turn there is a non-zero probability of misrecognizing and/or misunderstanding the user[1]. The system reacts to these errors by acting on it (e.g. using error recovery strategies) or ignoring them (e.g. asking the user to confirm the wrong intent). Second, there is a *compounding* effect on the user tendency to remain (with higher probability) in the negative state once it has been reached.

This behavior shows user tolerance to system errors is not time independent. DM strategies should take into account the current user state, emotion indicators as well as its history. As we will see in section 4, user state is a good predictor of system performance as well. Thus we might expect that a rising percentage of negative turns might be due to highly correlated variables such as user tolerance to system errors and to system over-reactions (e.g. inflated error-recovery sub-dialogs).

From Fig. 2 we infer how critical it is to engage the user into a positive state early on. Thus, we need to know also *when* ($\bar{t}$) to fire a specific DM action. In order to estimate the time $\bar{t}$ we have sampled a subset of all dialogs that contained at least one negative turn. We have then computed the probability that a negative state occurs at time $t = \bar{t}$ when preceding turns are *all positively* biased. In Fig. 3 we plot the estimate for the probability $P(t = \bar{t}|s(1) = 0, \ldots, s(\bar{t} - 1) = 0, s(\bar{t}) = 1)$.

From Fig. 3 we observe that, for those users that are bound to be in a negative state, the transition from positive to negative state will occur most likely early in the dialog. Such statistics could be exploited by the DM to calibrate the most

[1] The average WER is 27.7% and TCER is 16.3%

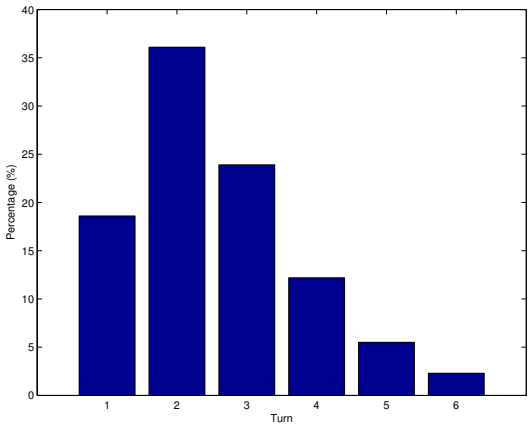

Fig. 3. Empirical estimate of time for transition probability ($P(t = \bar{t}|s(s(1) = 0, \ldots, s(\bar{t}-1) = 0, s(\bar{t}) = 1)$, of going into a negative user state, knowing that the user will change state in the next turn. The histogram is truncated at $t = 6$.

likely turn to fire strategies that are user state dependent. In the next section we will estimate the DM strategy distributions over the negative and positive labels and elaborate on state dependent DM actions.

3.2 Machine Actions

Another side effect of either poor system performance or inability to detect complete user state is the time it takes to complete the domain task. In Table 1 we give different statistics for the average length of the human-machine interactions. It is evident that the negative user responses will increase the dialog length by up to 30%. As the number of turns increase, it will lead to an increased probability of turning the user into a negative state (see Figure 2).

Table 1. The average dialog length in number of user turns for various conditions

	$p \geq 1$ $n = 0$	$n \geq 1$	$s(0) = 0$ $n \geq 1$	$s(0) = 0$ $s(t_f) = 1$
Dialog Length	3.2	4.4	4.6	4.4

Table 2 gives for each machine dialog act such as *Reprompt*, *Confirmation*, *Closing* and *Error Recovery* the distribution of negative user reactions. Most *Confirmation* moves occur when the system is confident (high intent posterior probability). *Closing* actions usually lead to either the end of the user dialog engagement or a transfer to a domain specialist or an automated system. Both *Confirmation* and *Closing* receive positive feedback from the user point of view. *Reprompt* and *Error*

Recovery are machine actions geared towards recovery of speech recognition and understanding errors. *Reprompt* actions are used at the very beginning of the interaction following the *"How may I help you?"* prompt. From Table 2 is evident that although most of the time the *reprompt* succeeds in maintaining the user in a positive state, almost 25% of the times it has a negative effect. A more compelling evidence of the negative effect is for two different *error recovery* strategies ((1) and (2) in Tab. 2). The two strategies differ for the prompt text realization and their usage over time. The relative frequency of occurrence turn numbers for each prompt is depicted in Figure 4. The second recovery strategy usually occurs later in the dialog and receives the largest negative responses among all DM actions. From Table 2 we observe that the second error recovery is penalized by achieving a poor feedback from the user.

4 Emotions and Machine Performance

In this section we quantify the impact of the user state on the performance of the spoken dialog both at the utterance level and in terms of the overall task success.

We randomly split the initial set of utterances into a training ($35K$) and a test set ($5K$). The test set has $1,344$ utterances labeled as having a negative user state, and $3,656$ utterances labeled as having a positive user state. We trained ASR models and SLU models on the training set annotated with transcriptions and 65 user intentions. For the ASR models we trained state-of-the-art acoustic and language models [6] and achieved test set WER of 27.7%. For the user call-types we trained a multi-class classifier based on the boosting algorithm [9]. We ran 1100 iterations and achieved an average of TCER 16.3%. In Table 3 we compute the word error rates for the two set of utterances with negative and positive emotion labels. There is a large gap in performance between the two classes (22% relative WER increase). This might be due to the indirect effect of

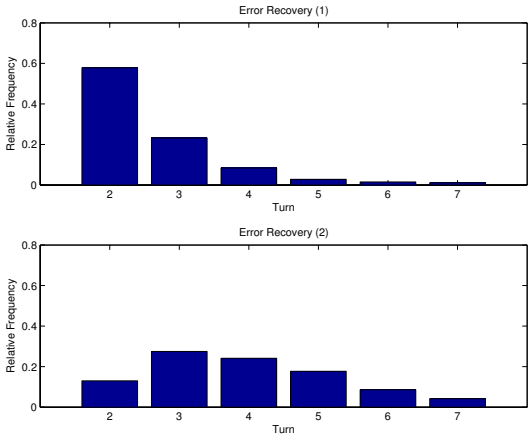

Fig. 4. Histograms (over time) for the two different Error Recovery strategies (1) and (2) in Tab. 2

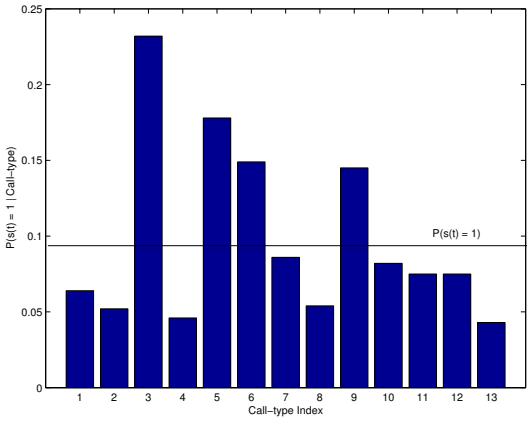

Fig. 5. Empirical estimate of the posterior probability $P(s(t) = 1|c_i)$ for call-type c_i (only the posterior probabilities of a subset of the 65 call-types is shown)

increasing the utterance length (see Table 2). We observe a similar pattern for the spoken language understanding task. This task is defined as the classification of user utterances at each turn, into one or more intent labels [4]. The increased classification error rate is due to the known limitations of SLU to handle long utterances [10]. These results support the finding that emotion predictors could be used to improve the prediction of word or classification error rate. Similarly, we might expect that the topic or intent of the user could be predictive of the user's emotional state. In Figure 5 we plot the posterior probability of having a negative user turn $P(s(t) = 1|c_i)$ for the most frequent call-types, c_i. Most of the semantic tag posterior probabilities fall below the prior probability $P(s(t) = 1)$, while a small set are significantly higher. The highest posterior corresponds to the *request for help*, as can be expected.

While WER and TCER provide an utterance-based system performance metric, dialog level metrics factor in the overall success of DM strategies in

Table 2. Utterance length (words) of spoken input, system performance in recognizing and understanding spoken utterances. Average error rates are computed for the negative and positive/neutral label partitions of the test set (Overall).

	Sentence Length (in words)	WER (%)	TCER (%)
Neutral State	7.9	24.8	14.7
Negative State	15.5	31.8	25.1
Overall	9.9	27.7	16.3

Table 3. The percentage of negative and positive states in response to various types of prompts (machine dialog acts)

	Positive State (%)	Negative State (%)
Reprompt	75.5	24.5
Error Recovery (1)	74.0	26.0
Error Recovery (2)	58.8	41.2
Confirmation	85.9	14.1
Closing	88.5	11.5

accomplishing the task. On the subset (647 dialogs) of the test set we have computed task success (failure) statistics and their association with different emotion traces. In Table 4 we show that there is a strong correlation between users consistently in positive state ($n = 0$) and task success (first column). Similarly, the final state (t_f) of the dialog's being negative ($s(t_f) = 1$) is a strong indicator of task failure. These statistics could be used to estimate prior conditions in the case of repeat-users. A sporadic transition into a negative state (second column, $n \geq 1$) does not necessarily correlate with the success (or failure) of the task completion. However, if the initial state of the user is positive and the user moves into a negative state, this is a strong indicator of task failure. The last two emotion trace statistics support the relevance of prior conditions in modeling the user state.

Table 4. Task success (failure) performance of the machine over different user state statistics

	$n = 0$	$n \geq 1$	$s(0) = 0$ $n \geq 1$	$s(0) = 0$ $s(t_f) = 1$
Task Success	70.7%	42.3%	38.9%	29.8%
Task Failure	29.3%	57.7%	61.1%	70.2%

5 Prior Conditions

Prior conditions refer to the user state being polarized negatively or positively prior to the user-machine interaction ($t = 1$). While the initial state might depend on a variety of causes directly or indirectly related to the actual user goal, it can greatly affect the expected user behavior and consequently impact the machine performance. In Fig. 2 we plot the histogram of negative state labels over time (turn) as the user-machine interaction proceeds within a dialog. At $t = 1$ the user is prompted with opening prompt (*How May I Help You?*) and state statistics show that 5% of users are negatively biased. The state dynamics

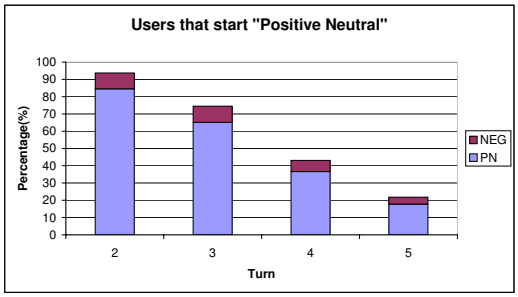

Fig. 6. Bar chart with percentage of negative/positive labels at each turn ($t \geq 2$) when $s(1) = 0$. The complement to 100% for each bin is the percentage of users that exit through the final dialog state or hang-up.

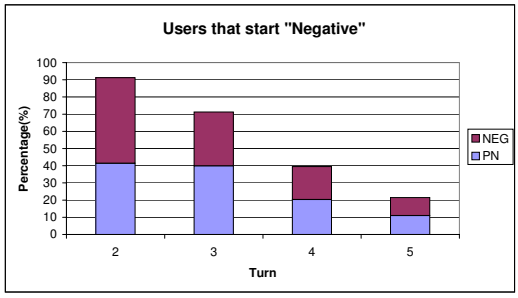

Fig. 7. Bar chart with percentage of negative/positive labels at each turn ($t \geq 2$) when $s(1) = 1$. The complement to 100% for each bin is the percentage of users that exit through the final dialog state or hang-up.

are significantly different for user groups with $s(1) = 0$ and $s(1) = 1$. In Figures 6 and 7 we plot a bar chart with percentage of negative and positive labels for $t > 1$ following a positive and negative initial turn, respectively. The complement to 100% for each bin is the percentage of users that exit through the final dialog state or hang-up. Fig. 7 shows that relative (with respect to positive) percentage of negative labels is constant in time. Therefore, if the user is in state $s(1) = 1$ it will be very unlikely (on average) to leave that state, given current system limitations. From this analysis it becomes evident how important it is to detect such prior conditions early in the dialog and adapt the machine's DM strategies accordingly.

6 Conclusion

In this paper we have investigated the role of emotions in human-machine spoken dialogs. Emotion levels have been quantized into positive/negative and user state statistics have been drawn from the *How May I Help You?* spoken dialog system.

For each human-machine interaction we have acquired the temporal emotion sequence going from the initial to the final conversational state. These statistical traces characterize the user state dynamics. We have grounded emotion patterns into dialog management strategies as well system performance. Our findings show that recognizing emotion temporal patterns can be beneficial to improve machine actions (dialog strategies) as well as to predict system error (ASR and SLU error rates).

Acknowledgements. We would like to thank Frederic Bechet for his contributions to the annotation of the dialog database.

References

1. Lee, C.M., Narayanan, S.: Towards detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing **13** (2005) 293–303
2. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proceedings of ICSLP, Denver, Colorado, USA (2002) 2037–2039
3. Litman, D., Forbes-Riley, K.: Predicting student emotions in computer-human tutoring dialogues. In: Proceedings of the 42nd Annual Meeting of the Association for Compuational Linguistics (ACL), Barcelona, Spain (2004)
4. Gorin, A.L., Riccardi, G., Wright, J.H.: How may I help you? Speech Communication **23** (1997) 113–127
5. Gupta, N., Tur, G., Hakkani-Tür, D., Bangalore, S., Riccardi, G., Rahim, M.: The AT&T spoken language understanding system. IEEE Transactions on Speech and Audio Processing (To appear)
6. Goffin, V., Allauzen, C., Bocchieri, E., Hakkani-Tür, D., Ljolje, A., Parthasarathy, S., Rahim, M., Riccardi, G., Saraclar, M.: The AT&T WATSON speech recognizer. In: Proceedings of IEEE ICASSP-2005, Philadelphia, PA, USA (2005)
7. Abella, A., Gorin, A.G.: Construct algebra: Analytical dialog management. In: Proceedings of the 42nd Annual Meeting of the Association for Compuational Linguistics (ACL), Washington D.C. (1999)
8. Shafran, I., Riley, M., Mohri, M.: Voice signatures. In: Proceedings of The 8th IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2003), St. Thomas, U.S. Virgin Islands (2003)
9. Schapire, R.E., Singer, Y.: BoosTexter: A boosting-based system for text categorization. Machine Learning **39** (2000) 135–168
10. Karahan, M., Hakkani-Tür, D., Riccardi, G., Tur, G.: Combining classifiers for spoken language understanding. In: Proceedings of IEEE workshop on Automatic Speech Recognition and Understanding, Virgin Islands, USA (2003)