# Recognition of Emotional States in Spoken Dialogue with a Robot

Kazunori Komatani, Ryosuke Ito, Tatsuya Kawahara, and Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
{komatani,kawahara,okuno}@i.kyoto-u.ac.jp

**Abstract.** For flexible interactions between a robot and humans, we address the issue of automatic recognition of human emotions during the interaction such as embarrassment, pleasure, and affinity. To construct classifiers of emotions, we used the dialogue data between a humanoid robot, Robovie, and children, which was collected with the WOZ (Wizard of Oz) method. Besides prosodic features extracted from a single utterance, characteristics specific to dialogues such as utterance intervals and differences with previous utterances were also used. We used the SVM (Support Vector Machine) as a classifier to recognize two temporary emotions such as embarrassment or pleasure, and the decision tree learning algorithm, C5.0, as a classifier to recognize persistent emotion, i.e. affinity. The accuracy of classification was 79% for embarrassment, 74% for pleasure, and 87% for affinity. The humanoid Robovie in which this emotion classification module was implemented demonstrated adaptive behaviors based on the emotions it recognized.

## 1 Introduction

A robot should be capable of interacting naturally with humans as a social partner and adapt its behavior according to his/her states. Emotions are important factors in reflecting these states [9], and therefore recognizing these plays an important role in dialogues particularly for entertainment. If a robot recognizes our emotions and responds in adaptation to these, we may feel social and friendly, which leads to more productive interaction.

Since speech interfaces play very important roles in human-robot interaction, automatic speech recognition (ASR) systems have recently been incorporated into robots for entertainment, such as pet and humanoid robots. Spoken dialogue technologies are also being introduced into them.

However, most recent research on spoken dialogue systems has only focused on verbal information contained in speech. Such systems, therefore, have tended to behave uniformly with all users when the verbal content of input sentences has been similar. Spoken dialogue, on the other hand, has many more characteristics than just verbal information. Such nonverbal characteristics also reflect individual user situations. The integration of nonverbal information should be taken into consideration to enable social interactions.

This paper focuses on emotional information, which has not been treated in conventional spoken dialogue systems. We present a method of automatically recognizing user's emotional states and achieving flexible dialogue based on emotions that can be recognized.

Most conventional studies into analyzing and recognizing speaker's emotions contained in speech have utilized prosodic features [2, 3, 8] and Kiebling et al. reported on these in detail [4]. We furthermore adopted another feature that is characteristics of dialogues, i.e. the interval between utterances. This is based on the assumption that this feature represents user's embarrassment.

We also addressed the issue of the classification without prior learning because we wanted to apply the method to robots interacting with unknown visitors. In general, a user's emotions included in speech are classified by comparing features in current utterances with those in his/her neutral states [7]. Therefore, data where a target user can be regarded as being in his/her neutral state is needed to normalize variations between individual users. We call the collection of data in their neutral states as prior learning. We designed several normalization methods that did not need prior learning, and attained comparable or better performance as a result.

There have been many classifications for human emotions such as anger, sadness, pleasure, calmness, surprise, and disgust. Huber et al. treated anger [2] and Lee et al. focused on negative emotions [6], to prevent customers on the telephone from hanging up. Our goal was to attain flexible interactions in a human-robot dialogue. We therefore focused on emotions that were important in spoken dialogue between humans and a robot, i.e., anger, pleasure, embarrassment, and affinity.

We evaluated our method using data collected from realistic situations. Many conventional studies have collected their data through having actors utter emotionally [7, 11]. We used data collected from children in a science museum with the WOZ (Wizard of Oz) method. The children's utterances were not pre-rehearsed but spontaneous. We also implemented our emotion recognition system in an interactive humanoid robot, Robovie [1], and achieved natural human-robot interactions.

## 2 Users' Mental States in Dialogues with Robots

We focused on emotions that were important in smoothing interactions between robots and humans. These emotions were derived after analyzing corpora that had been obtained from children interacting with a robot using the WOZ (Wizard of Oz) method. We specifically handled the following four emotions.

- **Anger**
  Users are often hurt by speech recognition errors, which are unavoidable in speech communications. Utterances when users are angry make speech recognition even more difficult. By detecting this emotion, the system assumes there has been some misunderstanding, and generates a response to relieve this.
- **Pleasure**
  If users look pleased, it is assumed that they are enjoying themselves, and the system does not need to change the topic. It then listens further on the topic.
- **Embarrassment**

If a user seems embarrassed about a topic, the system may change topics.

– **Affinity**
  There are many people who are not accustomed to talking with machines or robots. By detecting whether users are tense, the system can take action to alleviate this.

These emotions can be categorized into the following two according to their properties.

- **Temporary emotions**
  Temporary emotions vary per utterance, and affect the system's behavior during several utterances that follow. Anger, pleasure, and embarrassment can be categorized as temporary.
- **Persistent emotions**
  Persistent emotions depend on individual characteristics, and therefore do not change during one dialogue. The system's behavior is affected by the emotion throughout the whole dialogue. Affinity is categorized as persistent.

We then classified and evaluated temporary emotions per utterance and persistent emotions per speaker.

## 3 Target Data and Labeling

We used data that had been collected from visitors interacting with a robot using the WOZ method at the Kobe Science Museum. Most subjects were children aged from five to fifteen. The dialogue was in the form of questioning done by the robot on several topics. There were 46 dialogues, and a total of 498 utterances by subjects. The number of utterances ranged from ten and fifteen, and the duration of the dialogue lasted for several minutes. Figure 1 has a simple example dialogue.

Subjective labels were annotated for the data by three annotators, such as affinity (tense -- normal -- familiar), pleasure (pleased -- not pleased), and embarrassment (embarrassed -- not embarrassed). We did not treat anger because there was no data containing this emotion. This was because the data was collected with the WOZ method, and speech recognition errors had not occurred there. Annotators judged pleasure and embarrassment, which were temporary emotions, for every utterance. Affinity was determined for each dialogue. Table 1 lists the results of labeling for pleasure and embarrassment by the three annotators (A, B, and C in Table 1). Table 2

**Table 1.** Results of labeling for embarrassment and pleasure per utterance

| Annotator | Embarrassment | | Pleasure | |
|---|---|---|---|---|
| | Embarrassed | Not Embarrassed | Pleased | Not pleased |
| A | 81 | 417 | 193 | 305 |
| B | 67 | 431 | 72 | 426 |
| C | 76 | 422 | 168 | 330 |
| Agreement | 36 | 405 | 47 | 251 |

R: Hi.
U: Hi.
R: I am Robovie. Who are you?
U: I am Natsuko.
R: Oh, your name is Natsuko. How old are you?
U: Seven.
R: Seven years old. Are you a second grader?
U: Yes. Ha... How do you know that?
**(The subject is identified as pleased.)**
R: Do you like your school?
**(Asking the same topic further.)**
U: Yes.
R: What is interesting at the museum today?
U: Well ... That is ...
**(The subject is identified as embarrassed.)**
R: OK, I will change the topic. Do you like me?
**(Changing the topic.)**
U: Yes, I do.
R: Good!!

**Fig. 1.** Example dialogue between subjects and robot by WOZ (R: robot, U: user)

**Table 2.** Results of labeling for affinity per dialogue

| Affinity | | | |
|---|---|---|---|
| Familiar | Normal | Tense | Total |
| 17 | 21 | 8 | 46 |

is for affinity, which was determined by a single annotator for the 46 dialogues.

As we can see in Table 1, there were differences between the annotators. This means that impressions created by an utterance differ, and our aim was to recognize emotions that could be judged similar. We therefore used utterances for which the three annotators had given same labels in the experiment and evaluation that followed.

## 4 Automatic Classification of Users' Emotions

We now describe the automatic classification method of a user's emotions from his/her utterances. Although previous studies [7] needed prior learning where the characteristics of target users had been obtained beforehand, we used a method without the prior learning.

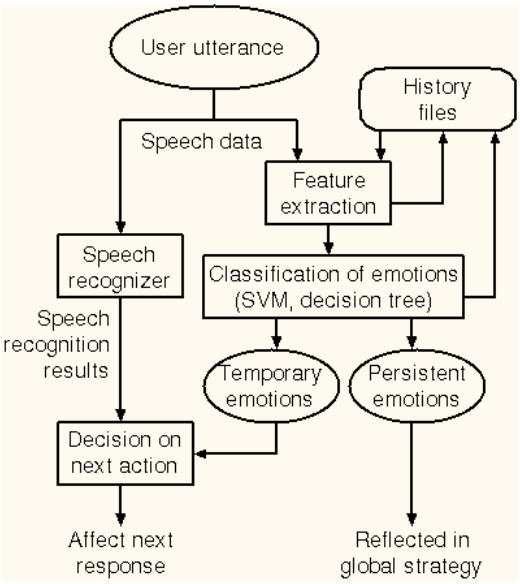


**Fig. 2.** Flowchart for system

Figure 2 is a flowchart for the system. It classifies a user's mental states from features contained in his/her speech. The classified mental states consist of temporary and persistent emotions as described in Section 2. User utterances are also transcribed by a speech recognizer. The next action is determined based on temporary emotions and the speech recognition results. The global strategy also changes based on the persistent emotion. Classification results and features of previous utterances are also taken into account as a history in addition to the features of the current utterance.

We used prosodic features that had been mentioned in many previous studies [2, 3, 4, 8], and the one we adopted can be listed as follows.

- Maximum F0 value
- Initial F0 value (onset)
- Average F0
- Difference between maximum and minimum F0
- Maximum power
- Average power
- Duration of utterance

Since we aimed at applying this method to robots interacting with unknown visitors, we could not normalize current utterances using values when users were calm. We therefore used another feature and normalization methods that were characteristic of dialogue, where the sequences of utterances were available. Specifically, we took the interval between the previous and current utterances, the difference of each feature with previous utterances, and normalization using the initial utterance of the dialogue into account.

- Difference between previous utterance and its normalization by current utterance
  The absolute values of features cannot be compared directly because they are individual. However, the differences between features are not affected by individuality, comparatively speaking. We considered both the differences and those normalized by features of the current utterance for each dimension.
- Normalization based on initial utterance
  We also normalized features of current utterances using those of initial utterances in the dialogue. This was based on the assumption that users at the beginning of the dialogue would be at their calmest.

We consequently adopted 29 feature values that consisted of the above seven features themselves and three operations (difference from previous utterance, difference normalized by current utterance, and normalization by initial utterance) for each of the seven features, and the interval between utterances. The interval was defined as the time between the end of the previous utterance and the beginning of the current utterance. We used the decision tree learning algorithm C5.0 [10] and the Support Vector Machine (SVM) [12] as classifiers. The linear kernel function was adopted for the SVM because there was not that much data.

## 5 Experiments and Evaluation

### 5.1 Experimental Conditions

We evaluated our method with the data described in Section 3, and the experiments were carried out with 10-fold cross validation. A process, where one tenth of all the data was used as the test data and the remainder was used as the training data, was repeated ten times, and the average accuracy was computed. We randomly changed the way the data was divided ten times, and computed the accuracy. The result obtained was averaged from a total of 100 calculations. The experiment for affinity was carried out by 5-fold cross validation because little data was available. To smooth out the unbalanced distribution of labels, we also introduced a cost corresponding to the reciprocal ratio of the number of samples in each class. This cost meant that the accuracy was computed under conditions where the number of samples was same for all classes.

### 5.2 Temporary Emotions (Embarrassment, Pleasure)

We will first describe the experimental evaluations we did on temporary emotions such as embarrassment and pleasure. As a baseline, we calculated the average values for utterances that were labeled as not having each emotion in the corpus, and normalized features of current utterances with this average. This baseline method

**Table 3.** Classification accuracy for embarrassment and pleasure (decision tree)

| Accuracy (%) | Embarrassment | Pleasure |
|---|---|---|
| Normalization by calm utterances (baseline) | 63.1 | 66.9 |
| Difference normalized by current utterances | 59.3 | 66.3 |
| Normalized by initial utterances | 66.3 | 68.0 |
| Using interval between utterances | 66.4 | 68.8 |
| Using all features | 69.0 | 66.8 |

**Table 4.** Classification accuracy for embarrassment and pleasure (SVM)

| Accuracy (%) | Embarrassment | Pleasure |
|---|---|---|
| Normalization by calm utterances (baseline) | 73.5 | 71.8 |
| Difference normalized by current utterances | 78.3 | 73.6 |
| Normalized by initial utterances | 75.4 | 72.9 |
| Using interval between utterances | 76.6 | 72.5 |
| Using all features | 79.0 | 71.9 |

corresponded to previous studies where features had been normalized by an utterance when users were calm that had been collected beforehand.

We calculated accuracy for the following conditions:

- Using both differences with the previous utterance and those normalized by the current utterance
- Using normalization based on the initial utterance
- Using the interval between utterances

We also calculated accuracy where all these 29 features were used.

Table 3 lists classification accuracy made by the decision tree trained by the C5.0 [10]. Classification accuracy was 69.0% for embarrassment, which was better than with the baseline method, which is equivalent to doing prior learning. The interval between utterances often appeared in the higher parts of the decision tree for embarrassment. This meant the feature was effective in classifying embarrassment, and was independent of the other features as it improved accuracy by being used together with them. There were no dominant features for pleasure.

Table 4 lists the classification accuracy obtained with the SVM [12]. We attained an accuracy of 79.0% for embarrassment and 73.6% for pleasure, which exceeded those with the baseline method. We analyzed significant features in classifying emotions by calculating accuracy where features were removed one by one. Features that played an important role in classifying embarrassment were maximum value of power, average F0, and intervals between utterances. The maximum value of power and its average were effective in classifying pleasure.

Since our method obtained higher accuracy than the baseline, which needed prior learning, the features and operations we propose are appropriate in classifying emotions without prior learning.

**Table 5.** Classification accuracy for affinity (C.5.0)

| | Accuracy for three classes (%) | Accuracy for two classes (%) |
|---|---|---|
| Average for first utterance | 44 | 66 |
| Average for first two utterances | 57 | 87 |
| Average for first three utterances | 56 | 79 |

### 5.3 Persistent Emotions (Affinity)

Let us now describe the experiment for affinity, which is a persistent emotion. Since a persistent emotion does not change greatly per utterance, we used the values for the seven prosodic features listed in Section 4 and the intervals between utterances.

A persistent emotion needs to be detected in the early stages of dialogue because it affects the global strategies that the system follows throughout the dialogue. We therefore calculated averages of the features for the first, first two, and first three utterances, and used them as features values in classification.

Classification was done with the decision tree (C5.0). We did not use the SVM because there was insufficient data for learning. The classification accuracy was calculated both for three classes (tense – normal – familiar) and for two classes: tense and others. This was because recognizing whether users were tense was more important in the dialogues.

Table 5 lists the classification accuracy for affinity. We attained an accuracy of 87% in classifying the two classes. The maximum value of power was effective in the classification.

## 6 Implementation in Robot

We installed the proposed classification module into a robot, Robovie, which was developed by ATR Intelligent Robotics and Communication Laboratories [1]. It was equipped with a number of sensors and movement mechanisms, and it communicated and interacted with humans through speech and gestures.

We used Julian [5], which had been developed at our laboratory, for speech recognition. It decoded human speech into transcriptions using a specified grammar as a linguistic constraint. To reduce speech recognition errors, we restricted the vocabulary by preparing grammars and dictionaries for each state corresponding to system questions. The questions were deliberated to narrow down next user's responses. Stationary noise was removed by spectral subtraction.

Temporary emotions were classified by the SVM for every utterance, and persistent emotions were classified by decision trees that were trained by the C5.0 through the initial two utterances in the dialogues. Figure 3 is a photograph of a conversation between Robovie and a human subject. Figure 4 has an example dialogue.

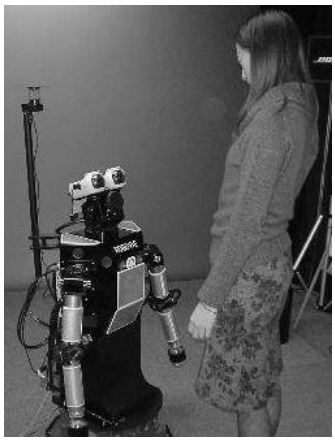

**Fig. 3.** Conversation with Robovie

R: Hello.
U: Hello.
R: Where are you from?
U: I come from Nagoya.
**(Change next action through detected emotions)**
R: Where is that?

**[In case of pleasure]**
R: Is it far or near? **(Pursue the topic)**
U: It is far from here.
R: You took a lot of trouble to come here, didn't you?
R: I come from ATR.

**[In case of embarrassment]**
R: I come from ATR.
**(Avoid pursuing the topic)**

… …

R: Do you think I am pretty?
U: Yes, you are.
**(Change next action by detected emotions)**
R: I am very glad to hear that.

**[In case of pleasure]**
R: Which part of mine is pretty?
**(Pursue the topic)**
U: All.
R: All? Yeah!

**[In case of embarrassment]**
R: Let's shake hands?
**(Change the topic)**
U: O.K.

**Fig. 4.** Example dialogue with proposed models

## 7 Conclusion

We addressed the issue of flexible interactions between robots and humans, and investigated emotions such as embarrassment, pleasure, and affinity. The emotions

were categorized into temporary emotions that changed per utterance and persistent emotions that did not change during dialogues. Conventional studies have needed prior learning to collect utterances in which users were calm. We, on the other hand, proposed the use of intervals between utterances and several operations that were specific to dialogue, such as calculating the differences between previous utterances and normalizing these using initial utterances. This enabled us to classify emotions without prior learning.

We also installed a classification module into a real robot. It changed its behavior according to emotions it recognized. Our future work will include evaluation of generated behaviors taking various experimental conditions and user characteristics into consideration.

## Acknowledgements

The authors are grateful to Professor Hiroshi Ishiguro and Dr. Takayuki Kanda of ATR-IRC for their help in installing the emotion classification module into Robovie.

## References

1. ATR Robovie. http://www.irc.atr.co.jp/~m-shiomi/Robovie/
2. Huber, R., Batliner, A., Buckow, J., Noth, E., Warnke, V., and Niemann, H.: Recognition of emotion in a realistic dialogue scenario. In *Proceedings of ICSLP* (2000), pp. 665-668.
3. Huber, R., Noth, E., Batliner, A., Warnke, V., and Niemann, H.: You BEEP Machine - Emotion in Automatic Speech Understanding System. In *Proceedings of TSD* (1998), pp. 223-228.
4. Kiebling, A., Kompe, R., Batliner, A., Niemann, H., and Noth, E.: Classification of Boundaries and Accents in Spontaneous Speech. In *Proceedings of the CRIM/FORWISS Workshop* (1996), pp. 104-113.
5. Lee, A., Kawahara, T., and Shikano, K.: Julius -- an open source real-time large vocabulary recognition engine. In *Proceedings of EUROSPEECH* (2001), pp. 1691-1694.
6. Lee, C. M., Narayanan, S. S., and Pieraccini, R.: Combining acoustic and language information for emotion recognition. In *Proceedings of ICSLP* (2002), pp. 873-876.
7. Moriyama, T., Saito, H., and Ozawa, S.: Evaluation of the Relationship between Emotional Concepts and Emotional Parameters on Speech. In *Proceedings of IEEE-ICASSP* (1997), Vol. 2, pp. 1431-1434.
8. Murray, I. R., and Amott, J. L.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of Acoustic Society of America*, Vol. 93, No. 2 (1993), pp. 1097-1108.
9. Picard, R. W.: Toward computers that recognize and respond to user emotion. *IBM Systems Journal*, Vol. 39, No. 3&4 (2000), pp. 705--719.
10. Quinlan, J. R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993. http://www.rulequest.com/see5-info.html
11. Schuller, B., Rigoll, G., and Lang, M.: Hidden Markov model-based speech emotion recognition. In *Proceedings of IEEE-ICASSP* (2003), Vol. 2, pp. 1-4.
12. Vapnik, V. N.: *Statistical Learning Theory*. John Wiley & Sons Inc., 1998.