Emotion Recognition by Combining Prosody with Text Information and Assessment Selection for Human-Robot Interaction

Yuanchao Li¹ Koji Inoue¹ Shizuka Nakamura¹ Katsuya Takanashi¹ Carlos Toshinori Ishi² Tatsuya Kawahara¹

¹ Graduate School of Informatics, Kyoto University
² Advanced Telecommunications Research Institute International (ATR)

Abstract: This paper addresses an emotion processing system, which consists of an emotion recognition module and an emotion generation module, for spoken dialog systems to improve human-robot interaction. In this study, two methods are proposed for realizing them. First, combining prosody with text information is performed for improving accuracy of emotion recognition. To utilize text information, a method named sentiment analysis is adopted. Moreover, emotion category and emotion level for assessment are predicted for the system to give human-like feedback. The distribution and occurrence frequency of emotion category are statistically analyzed, and the correlation coefficient between the dialog partners’ emotion levels is calculated.

1 Introduction

The progress of spoken dialog system technology and robot technology has made humanoid robots able to conduct more challenging tasks, including persuasion, guidance and counseling. On one hand, emotion became more and more important in human-robot interaction. To recognize emotion, two major types of models have been proposed and widely used. One is discrete models which insist on the existence of a small number of basic or fundamental emotions with specific eliciting conditions and response patterns in physiology and expression [1]. The other is dimensional models which map emotional states into a low-dimensional space [2]. Besides, some researchers found that combining text information could achieve better result [3].

On the other hand, making conversational agents give an occasional emotional feedback to the speaker became necessary. Verbal backchannels such as “okay” and “right” convey feedback. Without feedback, the speaker would be anxious whether the communication is well maintained. Backchannels suggest that the listener is listening, understanding, and agreeing to the speaker. Backchannels can also express listener’s feeling such as interest, surprise and sympathy [4]. The current systems still lack the ability to recognize nonverbal behaviors, especially they are insensitive to user’s feeling in their utterances and unable to prosodically show empathy.

This paper addresses an emotion processing system, which consists of an emotion recognition module and an emotion generation module, for spoken dialog systems to improve human-robot interaction. The contents of this paper are as follows. Chapter 2 describes the motivation of this research and the proposed system for solving limitations of the current spoken dialog systems. Chapter 3 explains the emotion recognition module by combining prosody and text information, and then shows the results. Chapter 4 explains the assessment selection function based on emotion

*連絡先：京都大学 情報学研究科
京都大学情報学専攻
京都市左京区吉田本町
E-mail: lyc@sap.iist.i.kyoto-u.ac.jp
generation module, which predicts both emotion category and emotion level, and then shows the analysis results. Chapter 6 concludes the paper.

2 System Overview

To make the android behave natural and human-like, emotion recognition and expression became necessary in spoken dialog systems. Hence, we propose a system which is shown in Figure 1. In this figure, the upper part is the spoken dialog system with the proposed system. The blue part is a simplified basic spoken dialog system and the green part is the proposed system.

ASR or Automatic Speech Recognition module transcribes spoken input into text in real time. Assessment Selection generates feedback with a rule-based procedure. The baseline spoken dialog system generates assessments only depending on the text generated by ASR, while the proposed system selects assessments depending on the user’s emotion. Assessment is a type of backchannel feedback, which is described Chapter 4. TTS or Text-To-Speech module converts output text into speech.

Emotion Processor is the proposed system, which takes prosody of human speech and text generated by ASR as input, and generates the system’s emotion as output. Then, the system selects an appropriate assessment based on the text and emotion.

The lower part of Figure 1 (green) is the flowchart of Emotion Processor. Left is the emotion recognition module (black), in which prosody and text are inputs. Right is the emotion generation module (red), in which emotion category and emotion level are collectively called system’s emotion which is the output of Emotion Processor.

This flowchart has five steps. The first step is from prosody to valence and arousal, the second from text to sentiment, the third utilizing sentiment, the fourth from valence and arousal to emotion category, and the last from valence and arousal to emotion level. Each step will be explained in the rest of this paper.

3 Emotion Recognition by Combining Prosody with Text Information

The study of speech emotion recognition has advanced greatly over recent years. In particular, it has become possible to infer user’s emotions from their voice thanks to models of acoustic and prosodic correlates of the various emotions. However, speech emotion recognition is still a challenging problem partly because it also depends much on text information. There are many researchers using text information in different ways to solve this problem [5]. Thus, we build an emotion recognition model by combining sentiment which is a kind of text information.

3.1 Corpus and Annotation

We choose the first meeting chat which consists of 15 dialog sessions between students and autonomous android Erica [6], which was remotely operated by a human operator. The corpus was collected by a Wizard of Oz manner. Figure 2 shows a scene of the dialog. Each dialog has two phases. In the first phase, Erica introduced herself and they talked about students’ lives, hobbies and futures. In the second phase, they talked about androids especially about Erica itself. The students were from the same university but different departments. Some of them had a little knowledge about androids and some did not. Each dialog session lasted around 15 minutes. We
3.2 Prosodic Features

We used Prosody Principal Components Analysis (PPCA) toolkit\(^1\) which supports prosodic analysis of speech and statistical methods to extract several useful prosodic features. The features we used are volume, creakiness, pitch lowness, pitch highness, narrow pitch, wide pitch and speaking rate. Each feature was computed over four time periods preceding the end point of each utterance. The time periods are -1600ms to -1100ms, -1100ms to -600ms, -600ms to -100ms, and -100ms to -0ms. In total, 28 features and PPCA’s standard normalization was used.

3.3 Modeling and Results

After labeling values and processing features, we conducted machine learning to find a predictor between emotion values and prosodic features. After mixing and shuffling the utterances, we chose 300 utterances, 250 for training and 50 for testing, to build the predictor. We used 50 test data to get 50 predicted values and used correlation coefficient for calculating correlation between the predicted values and the annotated values. We used 6-fold cross validation to make sure the model is robust. The average of correlation is 0.44 for valence and 0.59 for arousal. Correlation of 0.59 is good enough for building an emotion recognition model, but 0.44 is not satisfactory. We found that valence is often predicted incorrectly because it conflicts with sentiment, which is a subject feeling of text information. This is reasonable because Japanese people often say a negative fact with positive prosody or do not express positive feeling clearly even when they are happy.

3.4 Combination with Sentiment Analysis

Sentiment analysis refers to the use of natural language processing to identify and extract subjective information in source materials [7]. In this work, we used sentiment analysis to determine the affective state, specifically positive or negative feeling from the speaker’s utterance. We used Japanese Natural Language Processing \(^2\), a Python script which supports sentiment analysis of Japanese text.

\(^1\)http://cs.utep.edu/nigel/midlevel/mlv4.1.pdf  
\(^2\)http://jprocessing.readthedocs.io/en/latest/#

---

表1: An excerpt of annotation of the first meeting dialog corpus

<table>
<thead>
<tr>
<th>Subject</th>
<th>Transcription</th>
<th>Valence</th>
<th>Arousal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erica</td>
<td>あっ、まるごとにお仕事ですか。今、</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>あー、今はですね。えーと、豊中市。</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erica</td>
<td>あっ、なるほど、出身地はどちらですか。</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>出身は様島ですね。四国の。</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erica</td>
<td>常陸銀行。</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>萩島です。</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erica</td>
<td>あっ、様島病院、失礼いたしました。</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^*\) Disfluency of a slip of the tongue.

used four dialog sessions out of the fifteen sessions. Table 1 shows an excerpt of the corpus.

To describe emotion, we adopted a PAD model, which is a dimensional model\([6]\). For annotation of emotion, we prepared the following definition:

Valence: seven scales, from -3 (extremely negative) to +3 (extremely positive). This dimension represents the level of pleasure in the voice. Positive shows pleasant, whereas negative shows displeased.

Arousal: seven scales, from -3 (extremely passive) to +3 (extremely active). If a speaker is active, it sounds like he or she is engaged and shows high emotion in his or her voice. A passive voice would sound like a lack of engagement or low emotion.

We take only valence and arousal into consideration because of two reasons. First, our corpus is spontaneous speech instead of acted, and most speakers met in the first time. They kept friendly and did not show controlling and dominant feelings. Second, Japanese tend to speak in a calm and polite way because of the culture. So actually, the dominance dimension did not appear in our corpus.
We used two methods to combine text information. The first method is the weighted linear combination shown in formula (1). In this formula, $x$ is the valence prediction result from prosody, $y$ is the sentiment analysis result, and $\beta$ is the weight. We tested 0.5, 1, 2, and 3 for $\beta$.

$$z = x + \beta y$$

The second method is shown in formula (2). If the polarity of valence prediction result is the same as that of sentiment analysis result, we keep the value of valence recognition result. Otherwise, we change the polarity of valence prediction result.

$$z = \frac{xy}{x||y||}$$

Table 2 shows the average correlation coefficient of 6-fold cross validation and the statistical significance from the baseline result using prosody only. Table 3 shows some examples of valence prediction conflicts with the sentiment and the improved results by using the two methods. From Table 2, we can see that by combining the sentiment analysis, the correlation coefficient of valence is significantly improved, and the method 1 with $\beta = 3$ achieves the best result among all the conditions.

### 4 Assessment Selection

In recent years, a number of spoken dialog systems have been deployed in daily life to conduct task-oriented dialog. These systems basically conduct a transaction query, where the user makes one utterance per one turn, which is responded by the system. This communication mode differs a lot from human-human dialog, in which the listener occasionally gives feedback within the speaker’s turn. In this study, we focus on the assessment, which is backchannel feedback that often takes the form of words expressing degree of agreement, words of judgment, and words of sympathy and approval. We investigate a relationship in speaker’s emotion and listener’s emotion, which can be used to select an appropriate assessment.

### 4.1 Corpus and Annotation

We choose another first meeting chat which consists of seven dialog sessions between two human speakers. The topics they talked about include hobbies, fashion and news. Each dialog lasted 15 minutes to 25 minutes. This corpus was collected and annotated at ATR. We used four dialog sessions out of the seven sessions. Table 4 shows an excerpt of the corpus.

We annotated both the speaker’s emotion and the listener’s emotion. For the speaker’s emotion, we annotated valence and arousal of each utterance in the same manner as in Section 3.1. For the listener’s emotion, we prepared the following definition:

- **Category:** embarrassment, agitation, noticing, remembering, unexpectedness, surprise, puzzlement, anxiety, pain, dislike, disappointment, pleasure, anger
- **Level:** three scales, from 1 (slightly expressed) to 3 (extremely expressed)

Table 4 shows how we annotated. A is the speaker, B is the listener. We annotated the valence and arousal of A’s utterances, and the emotion category and emotion level of B’s utterances.
4.2 Assessment Category

An assessment feedback consists of both category and text. Since we only predict emotion, text is not our consideration. In Table 5, these categories are available with different levels of expression in the TTS ERICA system, and were designed based on previous works on backchannel analysis in human-human dialog [8]. In this chapter, we focus on a method to select an appropriate category and level of expression.

We divide these categories into three groups. The black ones are called attitude-related category, the blue ones are called partly emotion-related category, and the red ones are called emotion-related category. We focus on partly emotion-related and emotion-related categories because attitude-related category depends more on lexical information instead of emotion. In the rest of this paper, we call partly emotion-related and emotion-related categories collectively “emotion category” for convenience.

4.3 Emotion Category Prediction

In human-human dialog, the listener usually mimic the speaker’s emotion to build rapport. In this study, we focus on the relationship between two partners’ emotion categories. We annotated 240 utterance pairs and made distribution lists of emotion categories depending on valence and arousal together and separately. Then we analyzed the distribution lists aiming to find a general pattern. From Table 6 and Figure 3, we can see that when speaker’s valence is high, the listener tends to show a pleasure assessment. On the other hand, when speaker’s valence is low, the listener tends to show a disappointment assessment.

We can roughly predict that there is a high probability to use a pleasure emotion for high valence and a disappointment emotion for low valence. Although it is hard and not realistic to build a model to further select specific emotion category from this result, it could provide help for the future work. Hence, we let the system generate the two emotion categories at present.

4.4 Emotion Level Prediction

Emotion has not only category but also level, which should be decided when the system gives assessment. The same corpus and annotation were used in this chapter. For modeling, we first calculated the correla-
tion coefficient between each dimension and emotion level to see if they have good correlation. Then, we used machine learning to predict the system’s emotion level. After mixing and shuffling the data, we chose 200 for training and 40 for testing, to build the predictor using a linear regression as formula (1).

We used 40 test data to calculate a correlation coefficient between emotion levels of two dialog partners’. We also conducted 6-fold cross validation to make sure this model is robust. We added a minus sign to the level of negative emotions of disappointment, puzzlement and embarrassment.

The average correlation coefficient between the annotated level and the predicted level is 0.55. It is reasonable performance for predicting the system’s emotion level based on the user’s valence and arousal. The correlation coefficient between each dimension and emotion level is between 0.57 and 0.60. Both valence and arousal have good correlation with the emotion level.

5 Conclusion and Future Work

This study investigates an emotion processing system that consists of emotion recognition and assessment selection. Since the dimensional model performs better in dealing with spontaneous speech, we used it to estimate human user’s emotion. We introduced the sentiment analysis to improve emotion recognition by combining text information. The results show that valence prediction can be improved by using sentiment analysis. By predicting emotion category and emotion level, the system can select an appropriate assessment that makes human users feel comfortable and being understood. Given the condition that our current dialog system selects a prepared feedback which is based on the user’s utterance, we enhanced the system to make it emotional to express assessment. Hence, this method builds a bridge between the user’s dimensional emotion and the system’s discrete emotion.

The current emotion category prediction model can only determine positive or negative emotion. It cannot go further to select a specific emotion category. This is because the emotion depends on so many factors. In the future, we plan to take more factors into account, such as context and dialog act. In addition, we need more data to build a more general and robust system.

Acknowledgements

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

References


