Emotion Recognition by Combining Prosody and Sentiment Analysis for Expressing Reactive Emotion by Humanoid Robot

Yuanchao Li*, Carlos Toshinori Ishi†, Nigel Ward‡,
Koji Inoue*, Shizuka Nakamura*, Katsuya Takanashi*, Tatsuya Kawahara*
* Kyoto University, Kyoto, Japan
E-mail: {lyc, inoue, shizuka, takanasi}@sap.ist.i.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp
† ATR Hiroshi Ishiguro Labs., Kyoto, Japan
E-mail: carlos@atr.jp
‡ University of Texas at El Paso, El Paso, USA
E-mail: nigel@utep.edu

***Abstract*—In order to achieve rapport in human-robot interaction, it is important to express a reactive emotion that matches with the user's mental state. This paper addresses an emotion recognition method which combines prosody and sentiment analysis for the system to properly express reactive emotion. In the user emotion recognition module, valence estimation from prosodic features is combined with sentiment analysis of text information. Combining the two information sources significantly improved the valence estimation accuracy. In the reactive emotion expression module, the system's emotion category and level are predicted using the parameters estimated in the recognition module, based on distributions inferred from human-human dialog data. Subjective evaluation results show that the proposed method is effective for expressing human-like reactive emotion.**

I. Introduction

A number of spoken dialog systems have been investigated to conduct more challenging tasks, including persuasion, guidance and counseling. These systems basically focus on the accuracy of response generation instead of giving consideration to whether the users feel understood mentally. Expressing a reaction with proper emotion is necessary yet hard to realize partly because the systems are insensitive to the user's emotion, which makes them perceived as cold and robotic. To solve this issue, recognizing the user's emotion is the priority which has been researched for decades [1, 2, 3].

The study of speech emotion recognition has advanced greatly over recent years. In particular, it has become possible to infer users' emotion from their voice thanks to models of acoustic and prosodic correlates of the various emotions [4, 5], and two major types of models to describe emotion [6, 7]. However, speech emotion recognition is still a challenging problem because some emotions also depend much on text information. On the other hand, a reactive emotion such as surprise, agreement, sympathy, and approval is usually expressed to achieve rapport, by taking the form of backchannel feedback. Backchannel generation has been applied in some dialog systems, producing effective attentive listening behavior [8, 9]. However, it is not realistic to generate a specific reactive emotion due to the inaccuracy of the emotion recognition.

In this work, we aim at improving the emotion recognition, specifically valence, by combining prosody and sentiment analysis to make the system able to generate a proper reactive emotion. We first build a prosody-based emotion recognition model to analyze why valence is hard to estimate. Then, we incorporate sentiment analysis to solve the estimation errors. Finally, we predict the system's emotion category and level by using the improved emotion recognition results, for expressing the reactive emotion which is matched with the user's emotion. Subjective evaluation is conducted to verify the effectiveness of the proposed approach.

II. Analysis Data

A. Dialog Data for Emotion Recognition

Dialog data for emotion recognition were recorded with a humanoid ERICA [10]. During the data recordings, ERICA was remotely operated by a human operator in a Wizard of Oz manner. The subjects are students from the same university but different departments. Six sessions of the first meeting with ERICA were recorded. Each session consisted of 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. Each dialog session lasted around 15 minutes.

For annotation of the speaker's emotion, we prepared the following definitions:

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 unpleased.

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 adopted the utterances unit for labeling for convenience, taking an utterance to be a segment of speech that starts when a speaker begins a turn and ends when ERICA begins a turn.

The valence and arousal labels were annotated by two subjects. The agreement rate between the annotators was 77%. The disagreements were resolved through discussion between the annotators.

B. Dialog Data for Reactive Emotion Expression

We used another dialog data, which was recorded for the same project, consisting of seven dialog sessions between two human speakers who met each other for the first time. The topics they talked about include hobbies, fashion, and news, and were selected to contain topics of interest by both speakers or by only one of the counterparts. In such way, variations in the listener's reactive emotion expression are expected to occur more frequently. Each dialog lasted 15 minutes to 25 minutes.

We annotated both the speaker's emotion and the listener's emotion, in order to analyze the relationship between them. For the speaker's emotion, we annotated valence and arousal of each utterance in the same manner as in Section II.A. For annotation of the listener's emotion, we prepared the following definitions:

Category: embarrassment, unsettling, noticing, remembering, unexpectedness, surprise, hesitation, anxiety, pain, dislike, disappointment, pleasure, anger. These categories were designed by selecting emotion-related categories from past works on backchannel analysis [11, 12, 13].

Level of expression: three scales, from 1 (slightly expressed) to 3 (extremely expressed).

The emotion categories and levels were annotated by three subjects. The agreement rates by two or more annotators were 87% for category and 90% for level.

III. Emotion Recognition by Combining Prosody and Sentiment Analysis

A. Prosody-Based Emotion Recognition

We used Prosody Principal Components Analysis (PPCA) toolkit[1] which supports prosodic analysis of speech, and statistical methods to extract several useful prosodic features for machine learning works [14, 15, 16]. The features we used are energy, creakiness, pitch lowness, pitch highness, narrow pitch range, wide pitch range and speaking rate. Each feature was computed over four time periods preceding the end point of each utterance. The time periods are -1600 ms to -1100 ms, -1100 ms to -600 ms, -600 ms to -100 ms, and -100 ms to 0 ms. In total, 28 features and PPCA's standard normalization was used. Compared to the popular OpenSmile feature set[2], these features were designed to capture the dialog-relevant aspects of prosody.

After labeling values and processing features, we conducted linear regression to find a predictor between valence/arousal values and prosodic features. After shuffling the utterances, we chose 420 utterances and conducted 6-fold cross validation. The average correlation was 0.41 for valence and 0.62 for arousal. Correlation of 0.41 is not satisfactory for building an emotion recognition model. We found that valence is sometimes estimated incorrectly because it conflicts with sentiment, which is a subject feeling from text information. This is reasonable because people sometimes say a negative fact with positive prosody or do not express positive feeling clearly even when they are happy.

TABLE I
Results (correlation coefficient) of incorporating sentiment analysis to valence estimation.

Methods	Baseline (β = 0)	β = 0.5	β = 1	β = 2	β = 3
Average correlation coefficient	0.41	0.45	0.47	0.52	0.56
Statistical significance from baseline	/	>0.1	>0.05	<0.05	<0.05

B. Incorporating Sentiment Analysis

Sentiment analysis (also known as opinion mining) refers to the task of automatically determining the polarity of a piece of text, whether it is positive, negative, or neutral [17]. Sentiment analysis can be robust to ASR (Automatic Speech Recognition) errors [18].

In this work, we chose Japanese Natural Language Processing[3], a Python script which supports sentiment analysis of Japanese text. This function does sentiment analysis on Japanese text using word sense disambiguation based on a Japanese WordNet and an English SentiWordNet. We replaced the English SentiWordNet by the Japanese SentiWordNet so that it directly classifies on polarity score using the Japanese SentiWordNet. The results of this function can be [-1, 1].

We used the weighted linear combination shown in formula (1), to incorporate sentiment analysis result y to the prosody-based valence result x. We tested 0.5, 1, 2, and 3 for β, the weight of sentiment analysis result, to find the best one.

$$z = x + \beta y \tag{1}$$

C. Results and Discussion

Table I shows the average correlation coefficient of 6-fold cross validation and the statistical significance from the baseline result using prosody only. From this table, we can see that by combining the sentiment analysis, the correlation coefficient of valence is significantly improved, and the weighted linear combination with β = 3 achieves the best result among all the conditions. The correlation coefficient is increased by 0.15. Compared to previous works on valence estimation, the result of 0.56 outperforms the most common result, which is under 0.50 [19, 20, 21].

[1] http://cs.utep.edu/nigel/midlevel/mlv4.1.pdf

[2] http://audeering.com/technology/opensmile/

[3] http://jprocessing.readthedocs.io/en/latest

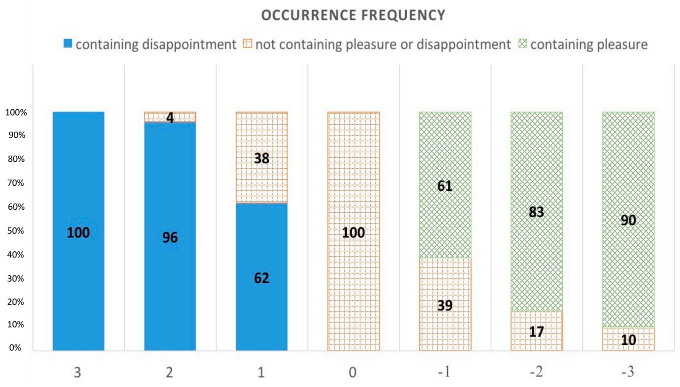

Fig. 1. Occurrence frequencies of the listener's emotion categories with and without pleasure or disappointment depending on the speaker's valence values.

We also tried the one-vector combination, which combines prosodic and textual features into a single vector when training the emotion recognition model. The improvement of this method is only 0.02, which is similar to previous works [19].

IV. Prediction for Reactive Emotion

We also propose a method to predict the system's emotion category and level based on the user's valence and arousal results obtained by using the proposed method introduced in Section III.

A. Emotion Category Prediction

We extracted 240 utterance pairs (speaker's statements and listener's reactive emotions) and analyzed the distributions of the listener's emotion categories depending on the speaker's valence and arousal, aiming to find a general pattern from the human-human dialog. Among the several categories (described in Section II.B), we firstly found that the categories "pleasure" and "disappointment" expressed by the listener had clear distributions with the speaker's valence.

We then classified the listener's utterances into three groups: utterances containing pleasure, utterances containing disappointment, and utterances not containing pleasure or disappointment. From Fig. 1, we can see that when the speaker's valence is positive, the listener tends to show his feeling containing pleasure emotion. The higher the valence is, the higher the occurrence frequency is. When the speaker's valence is low, the listener tends to show a disappointment emotion. The lower the valence is, the lower the occurrence frequency is. This phenomenon is reasonable because, in human-human communication, the listener usually shares the positive/negative feeling about a positive/negative fact with the speaker to express empathy.

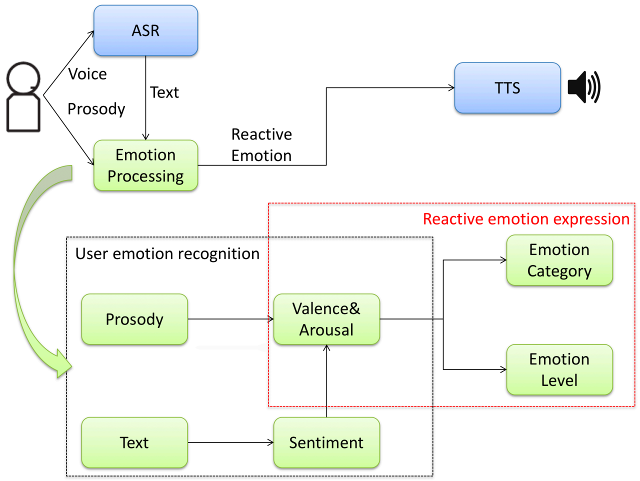

Fig. 2. Overview of the reactive emotion expression process.

B. Emotion Level Prediction

To build the emotion level prediction model, we first calculated the correlation coefficient between each dimension of the user's emotion and emotion level of the system to see if they have good correlation. Then, after shuffling the utterance pairs, we adopted linear regression and conducted 6-fold validation to predict the system's emotion level from the user's valence and arousal. We added a minus sign to the level of negative emotions of embarrassment, puzzlement, anxiety, pain, dislike, disappointment and anger.

The correlation coefficients between valence/arousal and emotion level were 0.57 and 0.60. Both valence and arousal have good correlation with the emotion level. This is reasonable because the two dialog partners are usually in similar levels of pleasure and activation. The average correlation coefficient between the annotated level and the predicted level was 0.55. We consider that this is a reasonable performance for predicting the system's emotion level based on the user's valence and arousal.

C. Overview of the Reactive Emotion Expression

The above-mentioned reactive emotion expression process is shown in Fig. 2. The system's emotion category and level are the outputs which are predicted based on the estimation of the user's valence and arousal.

V. Subjective Evaluation

We firstly prepared a text set of 20 sentences (10 with positive and 10 with negative contents) and 20 reactive emotion feedback utterances (such as "hontodesuka" "soudesune" and "naruhodo" which are equivalent to "really?" and "I see" in English). These reactive emotion utterances are available in our TTS system with different categories and levels of expression. The 20 sentences were recorded by 10 subjects (students aging from 21 to 24 years old), resulting in a total of 200 utterances. The subjects were only instructed to express their feelings corresponding to the sentences. Then, for each

TABLE II
SUBJECTIVE EVALUATION RESULTS FOR REACTIVE EMOTION EXPRESSION.

Methods	Are emotions proper?		Are feedbacks natural?	
	Mean	Variance	Mean	Variance
Neutral	0.0	1.1	-0.6	0.9
Random	-0.2	1.1	0.1	1.4
Proposed	1.4	0.3	1.1	0.5

utterance, valence and arousal values were estimated, and the system's emotion levels were predicted for expressing reactive emotion.

For comparison, we prepared the same lexical forms of reactive emotions using neutral and random emotion generation.

Neutral: The reactive emotions are generated in a neutral voice.

Random: The emotion category and level of reactive emotions are generated randomly.

Proposed: The emotion category and level of reactive emotions are predicted and generated based on the recognition results.

The same 10 subjects listened to three dialogs consisting of their own utterances recorded beforehand, followed by the reactive emotion feedback generated in the above-mentioned three conditions. Then they filled a questionnaire of two questions regarding their impressions of the system's reactive emotion feedback, on a five-point scale, from -2 to 2. The subjective evaluation results are given in Table II.

It is observed that the proposed method received better scores than the others. The results of the neutral condition suggest that the user may feel uncomfortable about most of the current dialog systems which do not take emotion into consideration. These systems cannot avoid the problem that their feedbacks are robotic and unnatural. Furthermore, we noted that when the subjects were listening to the dialog made by the proposed method, they nodded and smiled sometimes. The results along with this phenomenon suggest that making systems and humanoid robots emotional is needed.

VI. CONCLUSIONS

In this work, we have proposed an emotion recognition method that combines prosody and sentiment analysis for the system to properly express reactive emotion. First, we built a prosody-based emotion recognition model and verified that valence is hard to estimate from prosody because human speakers sometimes hold their feeling not showing the true emotion by voice. Then, we incorporated sentiment analysis to solve the problem that prosody-based valence estimation results sometimes conflict with sentiment analysis results. This method achieved a correlation improvement of 0.15, significantly outperforming the one-vector method which combines prosodic and textual features into a single vector. We also have proposed a reactive emotion expression method, where the system's emotion category and level are predicted using the parameters estimated by the emotion recognition. Subjective evaluation results showed that the proposed method generated emotion more properly and is effective for expressing natural and human-like reactive emotion.

ACKNOWLEDGMENT

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (Grant Number JPMJER1401), Japan.

REFERENCES

[1] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication,* vol. 40, no. 1, pp. 227256, 2003.

[2] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth, "Desperately seeking emotions or: Actors, wizards, and human beings," in *ISCA Tutorial and ResearchWorkshop (ITRW) on Speech and Emotion, 2000.*

[3] J. Pittermann, A. Pittermann, and W. Minker, *Handling emotions in human-computer dialogues.* Springer, 2010.

[4] A. B. Ingale and D. Chaudhari, "Speech emotion recognition," *International Journal of Soft Computing and Engineering (IJSCE),* vol. 2, no. 1, pp. 235238, 2012.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition,* vol. 44, no. 3, pp. 572587, 2011.

[6] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication,* vol. 40, no. 1, pp. 532, 2003.

[7] P. Ekman, "An argument for basic emotions," *Cognition & emotion,* vol. 6, no. 3-4, pp. 169200, 1992.

[8] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Ter Maat, G. McKeown, S. Pammi, M. Pantic et al., "Building autonomous sensitive artificial listeners," *IEEE Transactions on Affective Computing,* vol. 3, no. 2, pp. 165183, 2012.

[9] T. Kawahara, T. Yamaguchi, K. Inoue, K. Takanashi, and N.Ward, "Prediction and generation of backchannel form for attentive listening systems," in *Proc. INTERSPEECH,* vol. 2016, 2016.

[10] K. Inoue, P. Milhorat, D. Lala, T. Zhao, and T. Kawahara, "Talking with erica, an autonomous android," in *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue,* 2016, p. 212.

[11] C. T. Ishi, H. Ishiguro, and N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality," *Speech communication,* vol. 50, no. 6, pp. 531543, 2008.

[12] C. T. Ishi, H. Hatano, and N. Hagita, "Extraction of paralinguistic information carried by monosyllabic interjections in japanese," *Speech Prosody 2012,* pp. 681684, 2012.

[13] C. T. Ishi, H. Ishiguro, and N. Hagita, "Analysis of acousticprosodic features related to paralinguistic information carried by interjections in dialogue speech," *12th Annual conference of the International Speech Communication Association,* 2011, pp. 31333136.

[14] N. G. Ward and A. Vega, "Towards empirical dialog-state modeling and its use in language modeling." in *Interspeech,* 2012, pp. 23142317.

[15] N. G. Ward, S. D. Werner, F. Garcia, and E. Sanchis, "A prosody-based vector-space model of dialog activity for information retrieval," *Speech Communication,* vol. 68, pp. 8596, 2015.

[16] N. G. Ward and K. A. Richart-Ruiz, "Patterns of importance variation in spoken dialog," *14th SigDial,* 2013.

[17] S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," *Emotion Measurement,* pp. 201238, 2015.

[18] F. Mairesse, J. Polifroni, and G. Di Fabbrizio, "Can prosody inform sentiment analysis? experiments on short spoken reviews," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2012, pp. 50935096.

[19] M. Asgari, G. Kiss, J. Van Santen, I. Shafran, and X. Song, "Automatic measurement of affective valence and arousal in speech," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2014, pp. 965969.

[20] J. C. Acosta and N. G. Ward, "Achieving rapport with turn-byturn, user-responsive emotional coloring," *Speech Communication,* vol. 53, no. 9, pp. 11371148, 2011.

[21] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. (ICASSP).* IEEE, 2007, vol. 4, pp. IV1085.