A STUDY OF SOCIAL-AFFECTIVE COMMUNICATION: AUTOMATIC PREDICTION OF EMOTION TRIGGERS AND RESPONSES IN TELEVISION TALK SHOWS

Nurul Lubis, Sakriani Sakti, Graham Neubig, Koichiro Yoshino, Tomoki Toda, Satoshi Nakamura

Augmented Human Communication Laboratory
Graduate School of Information Science
Nara Institute of Science and Technology
{nurul.lubis.na4, ssakti, neubig, koichiro, tomoki, s-nakamura}@is.naist.jp

ABSTRACT

Advancements in spoken language technologies have allowed users to interact with computers in an increasingly natural manner. However, most conversational agents or dialogue systems are yet to consider emotional awareness in interaction. To consider emotion in these situations, social-affective knowledge in conversational agents is essential. In this paper, we present a study of the social-affective process in natural conversation from television talk shows. We analyze occurrences of emotion (emotional responses), and the events that elicit them (emotional triggers). We then utilize our analysis for prediction to model the ability of a dialogue system to decide an action and response in an affective interaction. This knowledge has great potential to incorporate emotion into human-computer interaction. Experiments in two languages, English and Indonesian, show that automatic prediction performance surpasses random guessing accuracy.

Index Terms: emotion, affective communication, dialogue system

1. INTRODUCTION

The study of social-affective communication is concerned with the role of emotion in human interaction. Social affective aspects of communication deeply enrich human interaction, highly affecting the way a speaker or listener behaves and responds. It is natural for humans to reflect their emotion in communication and be affected by their conversational counterpart. However, this is yet to be completely replicated in human-computer interaction (HCI).

The most widely researched sub-area of social-affective communication is *emotion recognition*, in which a computer attempts to recognize the emotion of a speaker from data in several modalities [1] [2] [3]. A number of emotion challenges have been held in the recent past, such as those at INTERSPEECH [4] [5] [6] as well as the Audio Visual Emotion Challenge (AVEC) [7] [8]. There has also been some work on *emotion simulation* in dialogue systems [9], where the computer attempts to give the impression that it is feeling a particular emotion.

In addition to these more traditional works on recognition and simulation, there has recently been an increasing interest in *emotion elicitation*, or *emotional triggers*, studying what causes emotion in the first place. Because emotion plays a two-way role in social-affective communication, knowing the reason behind displayed emotion is crucial in imitating conversations between humans. A recent study by Hasegawa et al. [10] addresses this issue by predicting and eliciting emotion in online conversation. We have followed up this study [11] by performing a similar task on the emotionally colored conversation between humans and simulated agents [12]. We have also performed a study expanding the largely English-centered previous work to Indonesian [13].

In this paper, we extend upon previous user-centered works that tried to predict and alter user's emotion. Instead, we try to predict a person's emotional reaction in a social-affective conversation by examining not only the person having the reaction, but also their conversational partner. To accomplish this, we analyze and observe how emotion fluctuates and how it is connected to actions taken in discourse. This analysis is intended to provide the knowledge to perform prediction of emotional triggers and responses. The prediction is aimed to accommodate the abilities required of a conversational agent or dialogue system to be emotionally aware: (1) to be able to decide an emotion triggering action, and (2) to be able to predict the appropriate response to an emotion trigger. Two languages are examined: English and Indonesian.

2. SOCIAL AFFECTIVE COMMUNICATION

In this study, we attempt to analyze social-affective aspects of an interaction and utilize them for prediction. To achieve this goal, we need a set of data that represents social-affective interaction. From this data, we would like to observe two major aspects of the conversation: (1) the emotional aspect, how emotion fluctuates and affects the conversation, and (2) the

978-1-4799-7291-3/15/$31.00 ©2015 IEEE

social aspect, concerning the conversational give and take between speakers. In this section, we describe the definition of properties of conversation to allow the aforementioned observations.

Defining and structuring emotion is essential in observing and analyzing its occurrence in conversation. We define the emotion scope based on the circumplex model of affect [14]. Two dimensions of emotion are defined: valence and arousal. Valence measures the positivity or negativity of emotion; e.g. the feeling of joy is indicated by positive valence while fear is negative. On the other hand, arousal measures the activity of emotion; e.g. depression is low in arousal (passive), while rage is high (active).

From the valence-arousal scale, we derive 4 common emotion terms: happiness, anger, sadness, and contentment. In correspondence to the valence-arousal dimensions, happiness is positive-active, anger is negative-active, sadness is negative-passive, and contentment is positive passive. Figure 1 illustrates these emotional dimensions.

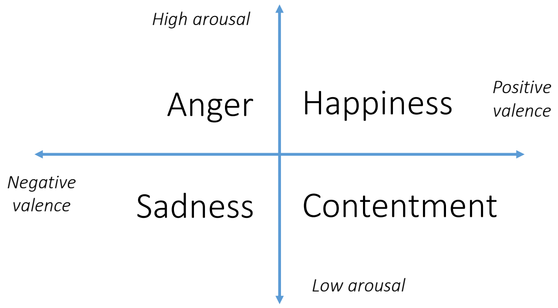


Fig. 1: Emotion classes and dimensions

On the other hand, to analyze the social aspect, it is necessary to observe the relationship of the utterances in a dialogue. To do this, we define a set of dialogue acts adapted from [15] to describe the structure of discourse. We reduce the original set of dialogue acts from 42 to 17 by grouping together similar acts, such as Yes-No-Question and Declarative Yes-No-Question. The 17 dialogue act labels are given in Table 1.

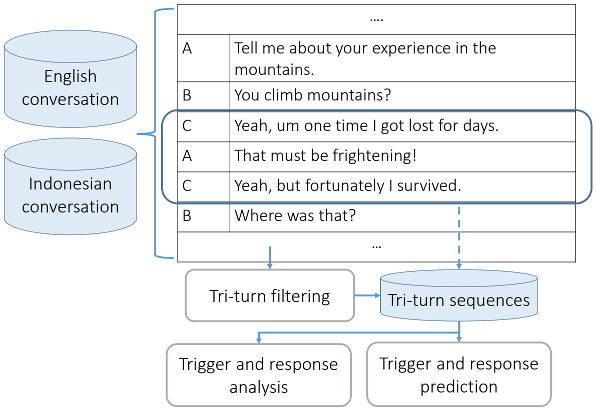
Another thing to pay attention to is that in natural conversation, interaction can be unordered and disconnected from one turn to the other. For example, an utterance can be abandoned, a speaker can get ignored, and the topic might drastically shift. Therefore, to properly analyze the fluctuation of emotion in a conversation, it is necessary to ensure that the observed sequences of conversation are in response to each other. To do this, we group consecutive sequences of conversation into a unit called a tri-turn [16]. Three consecutive sequences of speech in a conversation is considered a tri-turn when the second sequence is in response to the first, and the third is in response to the second.

Table 1: *Dialogue acts*

id	Dialogue Act	id	Dialogue Act
stat	Statement	rept	Repeat Phrase
opi	Opinion	ack	Acknowledgement
back	Backchannel	thnk	Thanking
Qyno	Yes-No Question	apcr	Appreciation
Qopn	Open Question	aplg	Apology
Qwh	Wh Question	hdg	Hedge
Qbck	Backchannel Question	drct	Directive
conf	Agree/Confirm	abdn	Abandoned
deny	Disagree/Deny		

The following sections will demonstrate how these properties help us in observing social-affective aspects in a conversation.

3. TASK DESCRIPTION



	....
A	Tell me about your experience in the mountains.
B	You climb mountains?
C	Yeah, um one time I got lost for days.
A	That must be frightening!
C	Yeah, but fortunately I survived.
B	Where was that?
	...

Fig. 2: *Overview*

Figure 2 presents the overview of our experiment. Tri-turns will be the basis of observation and analysis in this study. On each tri-turn, we analyze the change of emotion from the first sequence to the third, and correlate it to the second sequence that acts as the trigger. We categorize changes of emotional state into 7 events, described in Table 2.

To examine the connection between dialogue acts and changes of emotion occurring in conversation, we adapt the Term Frequency-Inverse Document Frequency (TF-IDF) formula to measure the importance of each dialogue act in triggering a certain emotion event. This formula is written as (1)

$$tfidf(d,t,T) = f(d,t) \times \log \frac{\{t \in T\}}{1 + \{t \in T : d \in t\}}, \quad (1)$$

where d is the dialogue act, t is the emotion event, and T is the collection of events. $f(d,t)$ denotes the raw frequency of d in t. This score is calculated for each dialogue act on each emotion event. This score can inform us if a particular dialogue act is characteristic of a particular emotion event.

Following the analysis, we perform automatic prediction of emotional triggers and responses. Two experiments are carried out. First, given the two sequences of interaction, we try to predict the emotional response that will occur. Second, given an emotional response, we try to predict characteristics of the trigger of that response. In our experiments, we use 90% of our data for training and test the model with the remaining 10%.

4. DATA CONSTRUCTION

In this paper, we collect our data from television talk shows, which contain real conversation with natural emotion occurrences. In television talk shows, the participants converse naturally about various emotion-provoking topics. The host directs the conversation to ensure that the discourse is structured as well as interesting. Due to these facts, this data is ideal for social-affective analysis. Furthermore, television talk shows provide clean speech recordings with distinguishable dialogue turns, as well as high quality speech data.

We expand the Indonesian Emotional Speech Corpus (IDESC) [17] with English conversational data from various television talk shows. Previously, we have collected Indonesian conversational data consisting of 1 hour, 34 minutes, and 49.7 seconds of speech. In this paper, we include English conversational data consisting of 1 hour, 2 minutes, and 19 seconds of speech from various television talk shows.

In both languages, we select three episodes from different kinds of talk shows to cover a broader range of emotion. In Indonesian, we select three episodes with discussion on politics, entertainment, and humanities. In English, we collect the data from three episodes of two of the most popular American television talk shows with discussion on entertainment, life experiences, and family struggles.

Table 2: *Types of emotion events*

Type	Description
emo_drop	positive emotion (happiness or contentment) changes to negative emotion (anger or sadness)
emo_raise	negative emotion (anger or sadness) changes to positive emotion (happiness or contentment)
act_drop	level of activation changes from high to low
act_raise	level of activation changes from low to high
val_drop	valence changes from positive to negative
val_raise	valence changes from negative to positive
cons	no significant change is observed

4.1. Annotation procedure

In annotating the corpus, we bear in mind that language and culture affect how emotion is perceived and expressed in an interaction. We carefully select 6 annotators for the task, 3 for each language. Every annotator is required to be (1) a native speaker of the language used in the show, and (2) knowledgeable of the culture in the interaction of the show. To ensure consistency, we have each annotator annotate the full corpus.

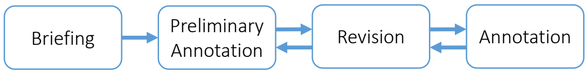


Fig. 3: Overview of annotation procedure

Figure 3 gives an overview of the annotation procedure. Before annotating the corpus, the annotators are briefed and given a document of guidelines to get a clearer picture of the task and its goal. The annotators proceed with preliminary annotation by working on a small subset of the corpus. This preliminary annotation lets the annotators get familiar with the task and confirms their annotation quality. After we screen the result, they are asked to revise inconsistencies with the guidelines if there are any. We perform the same screen-and-revise process on the full corpus annotation to achieve a tenable result.

4.2. Labels

With the axes in Figure 1, two sets of emotion labels are defined to allow observation from different perspectives. These sets are given in Table 3. The first emotion label set is for *emotion dimension*, consisting of the level of arousal and activation. The value of each dimension can be as low as -3 and as high as 3.

The second set is for *emotion class*; happiness, anger, sadness, and contentment. Instead of choosing which emotion is present, for each class, the annotators are instructed to rate its degree of presence. This rate ranges from 0 to 3, with 0 meaning that the emotion is not present and 3 meaning that the emotion is intensely present. This annotation scheme allows the observation of mixed emotion in speech.

To analyze the annotation consistency, we calculate mean Pearson's correlation coefficients *r* of the three annotators for each language, as shown in Figure 4. Moderate correlation is observed on all emotion labels except for contentment, which has weak correlation.

Dialogue act labels are defined according to Table 1. Annotators are asked to choose the proper dialogue act label on each dialogue turn. To analyze the consistency of dialogue annotation, we calculate Fleiss' kappa κ of the three annotators' results. Indonesian dialogue act annotation has a κ of 0.54 and English 0.45.

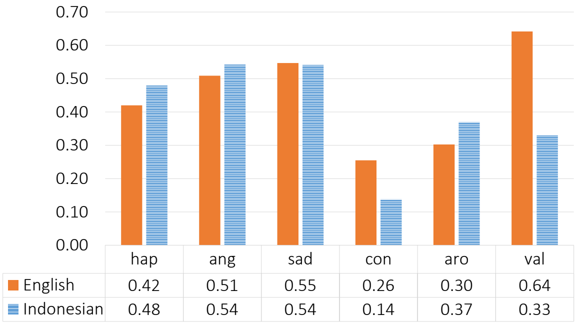


Fig. 4: Correlation coefficients of the emotion annotations

Table 3: *Emotion label sets*

id	Emotion Dimension	id	Emotion Class
aro	Arousal	hap	Happiness
val	Valence	ang	Anger
		sad	Sadness
		con	Contentment

5. ANALYSIS OF EMOTION DYNAMICS

We perform our analysis on the pre-processed data by correlating dialogue acts and the occuring emotion events. This section presents the result of the analysis and a potential application for emotional awareness in HCI.

5.1. Emotion dynamics and dialogue acts

To analyze the fluctuation of emotion in a conversation, on each tri-turn collected from our corpus, we observe the change of emotion from the first sequence to the third (emotional response), and correlate it to the second sequence (emotional trigger).

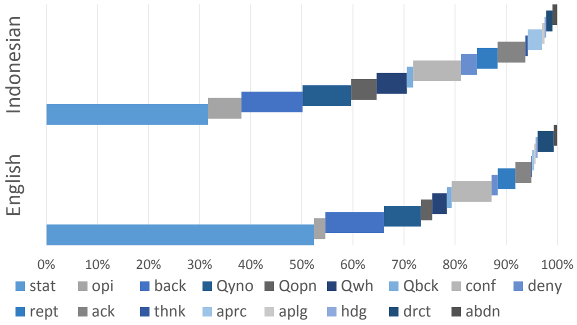
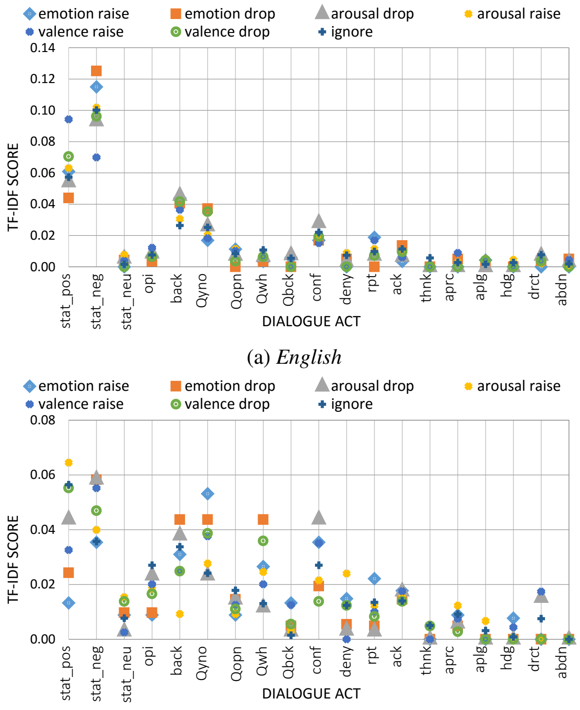


Fig. 5: *Dialogue act frequency on triggers of all emotion events*



(a) *English*

(b) *Indonesian*

Fig. 6: *Dialogue acts scores for all emotion events*

First, we take a look at all the emotional triggers, regardless of the emotion event they elicit, and see the frequency of dialogue acts. Figure 5 visualize the percentage of dialogue act frequencies for both languages. Statement has the highest frequency for both languages.

Second, we look into details for each emotion event. For each event, we calculate the TF-IDF score of each dialogue act using Equation (1). As the frequency of `stat` is high for all types of change, we separate statements according to their emotion, positive, negative, and neutral, and calculate their scores accordingly.

Figure 6 visualizes the TF-IDF scores of the dialogue acts, where a point in the graph corresponds to the TF-IDF score for a dialogue act on an emotion event. A large number of overlapping dots for a dialogue act means its TF-IDF score is similar for all emotion events. When this is observed, we can conclude that the dialogue act does not characterize a particular emotion event.

Comparison of Figures 6a and 6b tells us that the emotion events have different characteristics between the two languages when viewed from the action that triggers them. In English, the dots overlap one another with only slight differences, signaling that dialogue acts weakly characterize emotion events. On the other hand, in Indonesian, fewer overlapping dots are observed. This means, in English, emotion events are not attributed to the act taken in discourse, as opposed to Indonesian.

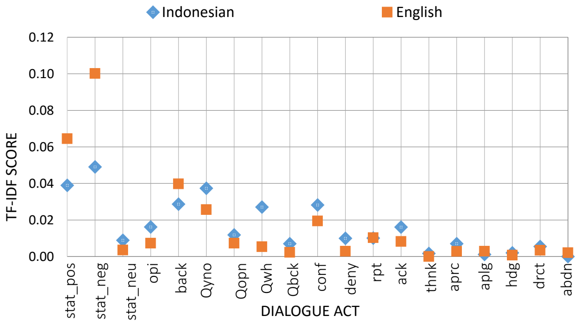


Fig. 7: *Average scores of dialogue acts in English and Indonesian*

Figure 7 shows the TF-IDF score of the dialogue acts in English and Indonesian averaged over all emotion events. From this figure, we can identify significant dialogue acts in triggering emotion events. In English, the top five dialogue acts are `stat`, `back`, `Qyno`, `conf`, and `ack`. In Indonesian, the top five dialogue acts are `stat`, `Qyno`, `back`, `conf`, and `Qwh`.

This finding suggests that showing interest in the conversation through questions and backchannels is a way to emotionally engage with the counterpart. Furthermore, providing new information in conversation can also elicit an emotional response from the counterpart. Table 4 includes an example of conversation taken from the corpus that demonstrates this finding. Looking at the level of valence and arousal, we can notice an increase for both the host and the guest.

Table 4: *Example of conversation*

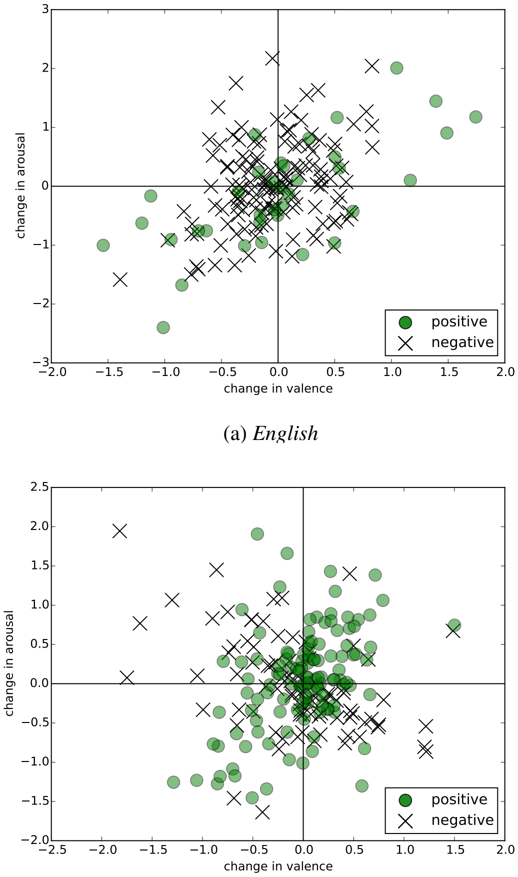
Speaker	Transcription	act	`aro`	`val`
Guest	Well I still have a lot of clothes in my closet I really shouldn't have	`stat`	0	-1
Host	Yeah	`back`	-1	0
Guest	But—yeah	`conf`	1	1
Host	Why?	`Qwh`	0	1
Guest	Just [inaudible] I want to say just in case but I don't think so 'cause I really think I got it conquered this time	`opi`	2	2

5.2. Emotion dynamics on statements

Statements make up a large part of human conversation. Therefore, we attempt to take a closer look at statements as emotional triggers by grouping them according to the emotion of the statement and plot them according to the emotional response. This distribution is shown in Figure 8.

Different tendencies can be observed from Figure 8. In English, the emotion triggering statements have an even distribution of positive and negative emotion. Statements with positive emotion spread to the upper right and lower left quadrants, meaning they cause an increase or decrease of both valence and arousal at the same time. On the other hand, statements with negative emotion give opposing effects to valence and arousal.

Unlike in English, in Indonesian, negative emotion dominated the emotion triggering statements. However, the few statements that have positive emotion seem to have a bigger effect compared to the negative ones.



(a) *English*

(b) *Indonesian*

Fig. 8: *Emotion of statements correlated to the emotion change it triggers*

6. AUTOMATIC PREDICTION OF EMOTIONAL TRIGGERS AND RESPONSES

In considering emotion, two of the important abilities of conversational agents or dialogue systems are (1) to be able to decide an emotion triggering action, and (2) to be able to predict the appropriate response to an emotion trigger. As the main goal of this study is incorporating emotion in HCI, we attempt to model the result of our analysis, allowing for automatic prediction to accommodate these abilities.

To gather information from the speech, we extract acoustic features as defined in the INTERSPEECH 2009 emotion chalenge [4] using the openSMILE feature extractor [18]. This feature set includes the most common yet promising feature types and functionals covering prosodic, spectral, and voice quality features.

On each tri-turn, we stack the features of the two corresponding tri-turn sequences to gather information of the context. To balance the number of instances and features, we perform correlation-based feature extraction [19] and linear discriminant analysis of our feature set. After reducing the dimension, we train a deep neural network classifier using Theano and the PDNN toolkit [20].

6.1. Automatic prediction of dialogue act of triggers

For dialogue act prediction, given the first and last sequence of a tri-turn as an emotional response, we try to automatically predict what action takes place as the trigger. We utilize the acoustic information, emotion annotation, and dialogue act of the first and last sequence as classification features. The prediction target is 17 dialogue acts presented in Table 1. This calculates to a chance rate of 5.88%.

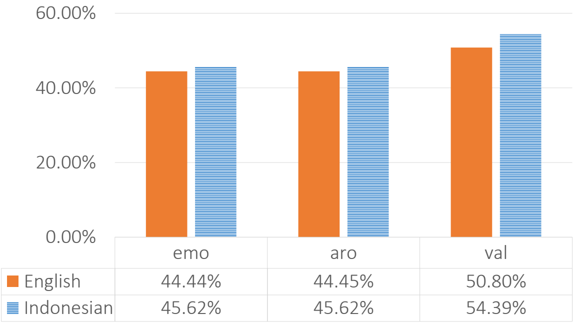
For the trigger prediction task, we achieve an accuracy of 53.97% for English, and 31.58% for Indonesian. The lower performance for Indonesian is suspected to be the implication of the dialogue act analysis in Section 5. In Indonesian, different emotion events can be triggered by different actions. Learning of these more complex patterns is likely to require a larger amount of data than currently present.

On the other hand, for English, triggers of a certain emotion event aren't strongly characterized by the dialogue act. In other words, the occuring emotion event only weakly affects the dialogue act, while other features such as acoustic features and dialogue act helps make the prediction.

6.2. Automatic prediction of emotional responses

Next, to simulate ability to understand social-affective dynamics in natural interaction, we attempt to predict the emotional response given an interaction of two speakers in a tri-turn. From tri-turn pairs in our corpus, we utilize the acoustic information, emotion annotation, and dialogue act of the first and second sequence as classification features.

We attempt to predict whether emotion events defined in Table 2 occur. As the emotion events can occur simultaneously, we examine the emotion class, valence level, and arousal level separately and try to predict whether `cons`, `raise`, or `drop` occurs. With three classes for prediction, we have a chance rate of 33.33%.



	emo	aro	val
🟧 English	44.44%	44.45%	50.80%
🟦 Indonesian	45.62%	45.62%	54.39%

Fig. 9: *Performance of response prediction*

Figure 9 presents our response prediction accuracy. For both languages, the prediction regarding valence has the highest performance, followed by arousal and emotion class. This could mean that valence and its fluctuations are expressed in speech more than is the case for arousal.

The suboptimal performance is likely due to the limited amount of data used in this study. Inherently, there are numerous factors that leads to a change of emotion in a conversation. This means that to properly recognize patterns for such events, a large number of features are required. It is likely that if we could prepare more data in the future, the accuracy will increase significantly.

7. CONCLUSIONS

In this paper, we presented a study on social-affective communication for automatic prediction of emotional triggers and responses. We examine natural conversations in English and Indonesian and analyze the emotion events that occurred within. Each language shows different tendencies and characteristics of emotional responses and triggers, suggesting that emotion events are language dependent. We look forward to affirm this finding, especially for English and Indonesian, on a larger amount of data.

In providing users the most natural HCI, emotion is an aspect that should not be overlooked. Our experiment on automatic prediction offers an approach in equipping conversational agents and dialogue systems with social-affective awareness. In future studies, we hope to include more modalities of interaction in observing the dynamics of emotion in interaction, such as textual and visual features. We also hope to incorporate this information directly into a dialogue system.

Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Number 24240032 and 26870371, as well as by a joint research project with Yanmar Co., Ltd.

8. REFERENCES

[1] W. Wang, G. Athanasopoulos, G. Patsis, V. Enescu, and H. Sahli, "Real-time emotion recognition from natural bodily expressions in child-robot interaction," in *Computer Vision-ECCV 2014 Workshops*. Springer, 2014, pp. 424–435.

[2] P. C. Petrantonakis and J. Leontios, "EEG-based emotion recognition using advanced signal processing techniques," *Emotion Recognition: A Pattern Analysis Approach*, pp. 269–293, 2014.

[3] K. R. Anne, S. Kuchibhotla, and H. D. Vankayalapati, "Emotion recognition using spectral features," in *Acoustic Modeling for Emotion Recognition*. Springer, 2015, pp. 17–26.

[4] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge." in *INTERSPEECH*, vol. 2009, 2009, pp. 312–315.

[5] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge." in *INTERSPEECH*, 2010, pp. 2794–2797.

[6] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," 2013.

[7] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011–the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 415–424.

[8] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

[9] C. Becker, S. Kopp, and I. Wachsmuth, "Simulating the emotion dynamics of a multimodal conversational agent," in *Affective Dialogue Systems*. Springer, 2004, pp. 154–165.

[10] T. Hasegawa, N. Kaji, N. Yoshinaga, and M. Toyoda, "Predicting and eliciting addressee's emotion in online dialogue." in *ACL (1)*, 2013, pp. 964–972.

[11] N. Lubis, S. Sakti, G. Neubig, T. Toda, A. Purwarianti, and S. Nakamura, "Emotion and its triggers in human spoken dialogue: Recognition and analysis," *Proc IWSDS*, 2014.

[12] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 5–17, 2012.

[13] N. Lubis, D. Lestari, S. Sakti, A. Purwarianti, and S. Nakamura, "Emotion recognition on Indonesian television talk shows," in *Proc IEEE Spoken Language Technology*, 2014.

[14] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[15] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[16] N. Lasguido, S. Sakti, G. Neubig, T. Tomoki, and S. Nakamura, "Utilizing human-to-human conversation examples for a multi domain chat-oriented dialog system," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 6, pp. 1497–1505, 2014.

[17] N. Lubis, D. Lestari, S. Sakti, A. Purwarianti, and S. Nakamura, "Construction and analysis of Indonesian emotional speech corpus," in *Proc Oriental COCOSDA*, 2014.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "OPENsmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[19] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[20] J. Bergstra, F. Bastien, O. Breuleux, P. Lamblin, R. Pascanu, O. Delalleau, G. Desjardins, D. Warde-Farley, I. Goodfellow, A. Bergeron *et al.*, "Theano: Deep learning on GPUs with python," in *NIPS 2011, BigLearning Workshop, Granada, Spain*, 2011.