Construction and Analysis of Social-Affective Interaction Corpus in English and Indonesian

Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura
Augmented Human Communication Laboratory
Graduate School of Information Science
Nara Institute of Science and Technology
{nurul.lubis.na4, ssakti, neubig, tomoki, s-nakamura}@is.naist.jp

***Abstract*—Social-affective aspects of interaction play a vital role in making human communication a rich and dynamic experience. Observation of complex emotional phenomena requires rich sets of labeled data of natural interaction. Although there has been an increase of interest in constructing corpora containing social interactions, there is still a lack of spontaneous and emotionally rich corpora. This paper presents a corpus of social-affective interactions in English and Indonesian, constructed from various television talk shows, containing natural conversations and real emotion occurrences. We carefully annotate the corpus in terms of emotion and discourse structure to allow for the aforementioned observation. The corpus is still in its early stage of development, yielding wide-ranging possibilities for future work.**

***Index Terms*—English, Indonesian, social-affective, emotion, corpus**

I. Introduction

Social-affective aspect concerns the incorporation of emotion in human interaction. This aspect is essential because emotion deeply enriches human communication, highly affecting the way a speaker behaves and responds in an interaction. It is natural for humans to reflect their emotion in communication and be affected by their conversational counterpart [1]. 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 [2] [3] [4]. However, because emotion plays a two-way role in social-affective communication, by itself, emotion recognition is insufficient in imitating conversations between humans. Given this problem, there has recently been an increasing interest in *emotion elicitation*, or *emotional triggers*, studying what causes emotion in the first place [5], [6].

Observation of complex phenomena of emotion requires rich sets of labeled data of natural interaction. Currently, many of emotionally colored speech corpora are constructed from acted speech or simulation [7] [8]. While this provides prominent emotion content, it does not represent natural occurrence of emotion. Furthermore, isolated emotional speech does not allow for the observation of emotion dynamics in an interaction. Although there has been an increase of interest in constructing corpora containing social interactions [9] [10], there is still a lack of spontaneous and emotionally rich corpora.

978-1-4673-8279-3/15/$31.00

We present a corpus of social-affective interactions in English and Indonesian. We construct the corpus from various television talk shows, containing natural conversations and real emotion occurrences. As the data are of similar nature and setup, the inclusion of two languages allows the observation of social-affective communication across different cultures. We annotate our data to provide rich information of the interaction while making sure of its quality and consistency.

This paper discusses about existing corpora in Sec. II. Sec. III describes how we collect the natural and emotionally rich interactions. The annotation procedure and labels are explained in detail in Sec. IV, followed by corpus analysis in Sec. V. Finally, we conclude this paper and discuss the future works in Sec. VI.

II. Related Works

There are the a number of difficulties that we have to overcome in constructing an appropriate emotional speech corpus. Emotion is inherently a very personal human experience and tends to be kept private. Furthermore, it is not something that can be replicated easily. This makes the collection of spontaneous, natural emotional speech for research purposes a sensitive matter, often raising moral and ethical issues [11]. In addressing these difficulties, several methods and approaches have been inspected in previous studies; e.g. emotion portrayal, induction, and simulation.

The Geneva Multimodal Emotion Portrayals (GEMEP) corpus was constructed from acted portrayals of emotion [7]. While these corpora do provide emotionally rich content, the usage of acted affect as a base of study or experiment raises a number of concerns due to the characteristics of acted emotion; it is stereotypical, lacking context, and limited to general and basic emotions. However, a previous study had argued that if designed carefully, acted emotion portrayals can nonetheless be beneficial in emotion research [12].

The SEMAINE Database consists of natural dialogue between user and operator simulating a Sensitive Artificial Listener (SAL) [8] with different traits. These different characteristics of the SALs elicit different emotions from the user, thus resulting in emotionally colorful data. The occurrence of

elicited emotion is more natural than acted emotion. However, the set-up sometimes gives unpredictable result as the participants are fully aware of the intention of the agents.

Other non-acted emotion corpora include the Vera am Mittag corpus, collected from German television broadcast [13], and HUMAINE Database, a multimodal corpus consisting of natural and induced data showing emotion in a range of contexts [9]. The usage of interview recordings is beneficial as they are natural and the emotion occurrences are more realistic.

In this paper, we collect our data from television talk shows, containing real conversation with natural emotion occurrences. In television talk shows, the participants converse naturally about various topics of discussion. The host directs the conversation to ensure that the discourse is structured, and to keep the content interesting for the listener by engaging with their guests and triggering utterances. This provides good social-affective data for analysis. Furthermore, television talk shows provide clean speech recordings with distinguishable dialogue turns, as well as high quality speech data.

III. Data Collection

As the corpus aims to capture realistic emotion that is applicable to human-computer interaction, we construct our corpus from a number of talk shows in both English and Indonesian covering various topics of discussion such as life experiences and politics, containing natural conversations and real emotion occurrences.

We expand the Indonesian Emotional Speech Corpus (IDESC) [14] 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. By considering two languages, we further the potential utilization of the corpus; e.g. for observing social-affective communication across different cultures.

In Indonesian, we select three episodes from different kinds of talk shows to cover a broader range of emotion. The selected talk shows are very popular in the country, with discussions on engaging and interesting topics that trigger various emotions from the speakers. The first show is contains discussion focusing on politic related subjects. The second show concerns with topics in the area of humanities. The third show is has a lighter focus in celebrities, their career, and life.

In English, we collect the data from three episodes of two of the most popular American television talk shows world wide. One of the shows talks about life experience of public figures, while the rest contain discussion of the struggles of families and negotiation in overcoming issues. In each language, the different topics are expected to provide varied emotion content in the collected data.

Video recordings of the show are obtained, but stripped down to audio only as we are currently focusing on speech data. Audio is available at 16 kHz and 16 bits per sample. In English, there are 12 speakers in total; 4 male speakers and 8 female. In Indonesian, there are 18 speakers in total; 12 male speakers and 6 female.

IV. Annotation

In this section, we explain the annotation procedure, where we impose rigorous quality control to ensure the consistency of the results. We also define and describe the annotation labels. The corpus provides emotion and dialogue act annotations of social-affective interaction. Furthermore, speaker information and speech transcription is also provided.

A. 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. With these requirements, we try to ensure that the annotators can observe emotion dynamics of the interaction to the furthest extent. To ensure consistency, we have each annotator annotate the full corpus.

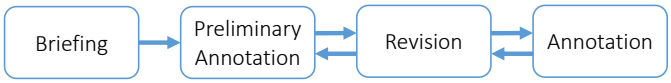


Fig. 1. Overview of annotation procedure

Fig. 1 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 document provides theoretical background of emotion and dialogue acts in discourse as well as a number of examples.

After the annotators are briefed, firstly, we ask them to do preliminary annotation by working on a small subset of the corpus. This step is done to let them get familiar with the task. Furthermore, with the preliminary result, we are able to confirm whether the annotators have fully understood the guidelines, and verify the quality and consistency of their annotations.

We manually screen the preliminary annotation result and give feedback to the annotators accordingly. They are asked to revise inconsistencies with the guidelines if there are any. This process is repeated until the quality of the preliminary annotation is sufficient. Once their results are verified, the annotators are authorized to work on the rest of the corpus. We perform the same screen-and-revise process on the full corpus annotation to achieve a tenable result.

B. Dialogue Act Annotation

As the corpus tries to capture social-affective interaction, we perform the annotation on the speech-turn level, from here onwards referred to as a segment, by determining its dialogue act. A dialogue act represents the meaning of an utterance at the level of illocutionary force [15]. The dialogue

act annotation will provide information about the discourse structure, showing the relationship between speech segments.

We define a set of dialogue acts adapted from [16] to describe the structure of discourse. We reduce the original set of labels from 42 to 17 by grouping together similar labels, such as Yes-No-Question and Declarative Yes-No-Question. The 17 dialogue act labels are given in Table I.

TABLE I
Dialogue act labels

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		

C. Emotion Annotation

Defining and structuring emotion is essential in observing and analyzing its occurrence in conversation. We define the emotion labels based on the circumplex model of affect [17]. 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. Fig. 2 illustrates these emotional dimensions.

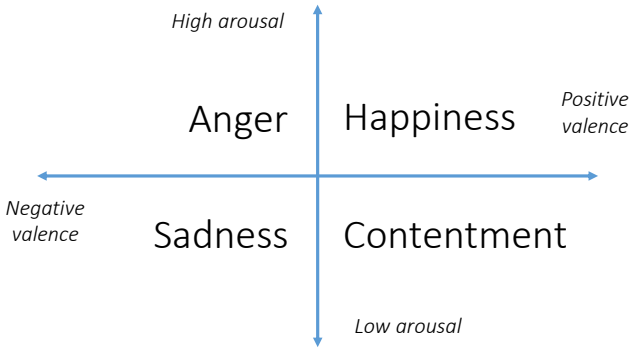


Fig. 2. Emotion classes and dimensions

With these axes, two sets of emotion labels are defined to allow observation from different perspectives. These sets are given in Table II. 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.

TABLE II
Emotion label sets

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

V. Analysis

We inspect three properties of the corpus to gain better insight of the data contained within. First, we take a look at the distribution of segment length. Secondly, we look over the conversational aspect of each language through the composition of dialogue acts in the corpus. Lastly, we examine the quality of emotion annotation by looking at the inter-annotator label correlation.

It is important to keep in mind that these analyses do not provide conclusive differentiation between English and Indonesian languages due to the limited amount of data at hand and the differences in conversation topics between the two languages. However, they may give some idea about the different phenomenon and tendency that occur between the collected American-English and Indonesian television talk show broadcasts.

A. Length of Segments

We plot the distribution of number of segments according to their duration on Fig. 3. In the figure, the y-axis shows the number of segments with respect to the x-axis, which shows the duration. Different trends can be observed in each language. The line graph for the English talk shows exceeds that of Indonesian for shorter durations, and then decreases heavily and has the value 0 from 16 seconds throughout the end. On the other hand, a long tail can be seen in the line graph for Indonesian.

The plot on Fig. 3 shows that English speech segments are shorter on average compared to that of Indonesian. Respectively, the average durations of segments for English and Indonesian are 2.93 and 7.57 seconds. This indicates that a dialogue turn in Indonesian tend to last longer than that in English. This difference of duration potentially affects the way emotion is communicated in a conversation.

B. Dialogue Act Labels

To analyze the consistency of annotation, we calculate Fleiss' kappa κ of the three annotators' results. κ measures

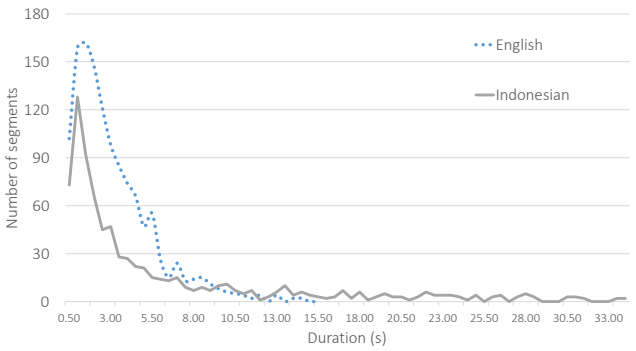
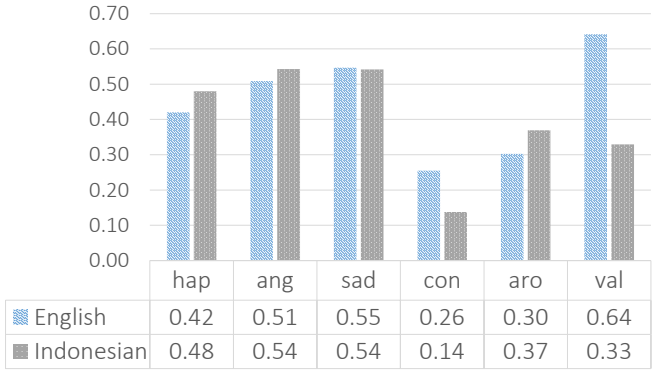


Fig. 3. Number of segments in respect to the duration



	hap	ang	sad	con	aro	val
English	0.42	0.51	0.55	0.26	0.30	0.64
Indonesian	0.48	0.54	0.54	0.14	0.37	0.33

Fig. 5. Correlation coefficients of the emotion annotations

the inter-annotator agreement of nominal variables when more than two annotators are employed. Respectively, English and Indonesian dialogue act annotation have κ of 0.54 and 0.45.

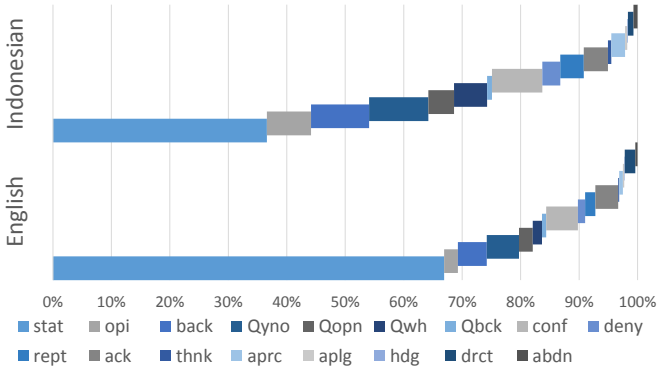


Fig. 4. Composition of dialogue act labels

Fig. 4 shows the composition of dialogue acts in the social-affective interaction on television talk shows. Statements dominate the collected conversations in both languages. This is not a surprising finding, as information exchange often happens in the form of a statement. However, it can be observed that in Indonesian, the composition is less dominated with statements than English. This could indicate more social-affective feedback and activity in the interaction, for example, in form of questions, confirmation, and back channel.

C. Emotion Labels

To analyze the annotation consistency, we calculate mean Pearson's correlation coefficients r of the three annotators for each language. Pearson's r measures the strength and direction of linear relationship between two variables. An absolute value of r between 0.0 and 0.3 is interpreted as weak correlation, and greater than 0.3 up to 0.5 as moderate correlation. Moderate correlation is observed on all emotion labels except for contentment, which has weak correlation.

Fig. 5 presents the correlation coefficient of the emotion annotation. The difficulty of identifying contentment in speech seems to occur in both languages. This is probably due to the subtleness of passive-positive emotion, thus sometimes can be overlooked. In general, the rates of correlation for the labels seem to be comparable for both language. However, valence, the second least correlated label in Indonesian, instead has the best rate in English. On average, the inter-annotator correlation for English is 0.44, and 0.46 for Indonesian; both can be interpreted as moderately correlated.

VI. Conclusion and Future Works

We presented a corpus of social-affective interactions in English and Indonesian. We constructed the corpus from various television talk shows, containing natural interaction with spontaneous emotion occurrences. As the data were collected from similar sources, the inclusion of two languages allows the observation of social-affective communication across different cultures. We carefully annotated our data to provide rich information of the interaction while making sure of its quality and consistency.

Increasing the amount of data is certainly a way to improve the value of the corpus. With more data, the corpus will contain more emotion variation and interaction dynamics. Aside from that, the inter-rater agreement shall be resolved and improved; for example, by using an emotion-tracing tool such as [18], and employing more annotators.

In the future, we intend to carry a more in-depth study of social-affective communication with the corpus. In doing so, we would consider acoustic and lexical features to arrive at a more meaningful findings.

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.

REFERENCES

[1] V. Christophe and B. Rimé, "Exposure to the social sharing of emotion: Emotional impact, listener responses and secondary social sharing," *European Journal of Social Psychology*, vol. 27, no. 1, pp. 37–54, 1997.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] T. Bänziger, H. Pirker, and K. Scherer, "Gemep-geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions," in *Proceedings of LREC*, vol. 6, 2006, pp. 15–019.

[8] 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.

[9] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: addressing the collection and annotation of naturalistic and induced emotional data," in *Affective computing and intelligent interaction*. Springer, 2007, pp. 488–500.

[10] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[11] 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. 572–587, 2011.

[12] T. Bänziger and K. R. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus," in *Affective computing and intelligent interaction*. Springer, 2007, pp. 476–487.

[13] M. Grimm, K. Kroschel, and S. Narayan, "The vera am mittag german audio-visual emotional speech database," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2008.

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

[15] J. L. Austin, *How to do things with words*. Oxford university press, 1975, vol. 367.

[16] 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.

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

[18] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.