Construction of Japanese Audio-Visual Emotion Database and Its Application in Emotion Recognition

Nurul Lubis[†], **Randy Gomez**[‡], **Sakriani Sakti**[†], **Keisuke Nakamura**[‡], **Koichiro Yoshino**[†], **Satoshi Nakamura**[†], **and Kazuhiro Nakadai**[‡]
[†]Nara Institute of Science and Technology, [‡]Honda Research Institute Japan Co., Ltd.
[†]Ikoma, Nara, Japan; [‡]Wako, Saitama, Japan
[†] {nurul.lubis.na4, ssakti, koichiro, s-nakamura}@is.naist.jp,
[‡] {r.gomez, keisuke, nakadai}@jp.honda-ri.com

Abstract

Emotional aspects play a vital role in making human communication a rich and dynamic experience. As we introduce more automated system in our daily lives, it becomes increasingly important to incorporate emotion to provide as natural an interaction as possible. To achieve said incorporation, rich sets of labeled emotional data is prerequisite. However, in Japanese, existing emotion database is still limited to unimodal and bimodal corpora. Since emotion is not only expressed through speech, but also visually at the same time, it is essential to include multiple modalities in an observation. In this paper, we present the first audio-visual emotion corpora in Japanese, collected from 14 native speakers. The corpus contains 100 minutes of annotated and transcribed material. We performed preliminary emotion recognition experiments on the corpus and achieved an accuracy of 61.42% for five classes of emotion.

Keywords: Japanese, emotion, corpus, audio-visual, multimodal

1. Introduction

Social interaction between humans are highly colored with emotion occurrences — we console a sad friend, share a happy experience, and debate on a disagreement. It is argued that humans also impose this emotional aspect in their interaction with computers and machines (Reeves and Nass, 1996). They treat them politely, laugh with them, and sometimes get angry or frustrated at them. In that case, to achieve a natural and human-like interaction, it is critical that computers be capable to reciprocate in these affective patterns.

Particularly in Japan, the need of an emotion sensitive system continues to escalate. The big number of depression cases, the trend of an unhealthy working habit (Kitanaka, 2011), and the aging population are among many conditions where assistive technology is in serious need. A system supporting intensive care and treatment will be of valuable aid in addressing these issues. In all of these cases, affective aspects are undoubtedly fundamental.

To achieve an emotion sensitive system, various capabilities are required; the system has to be able to recognize emotion, taking it into account in performing its main task, and incorporate it in interacting with the user. Many research and studies have been done globally regarding these issues (Wang et al., 2014; Petrantonakis and Leontios, 2014; Iida et al., 2003; Bulut et al., 2002). In all of these efforts, emotion corpus is a prerequisite.

In Japanese, emotion corpora are still limited to textual, speech (Arimoto et al., 2008), facial expression (Lyons et al., 1998), and physiological signals (Zhang et al., 2014) separately. Given that humans express emotion through multiple channels, a multimodal corpus is important to present a more complete information of emotion occurrences. In this paper, we present an audio-visual emotion corpus in Japanese, containing various emotion portrayals from 14 native speakers. Our approach is novel in that we provide contextualized emotion portrayals in addition to the isolated ones by using both monologues and situated dialogues in the emotion portrayal. In total, the corpus contains approximately 100 minutes of annotated and transcribed material.

Sec. 2. discusses existing corpora and their approaches. We define the scope of emotion within this paper in Sec. 3. The data collection procedure is described in Sec. 4., followed by an explanation of the annotation procedure and labels in Sec. 5. We present the result of a preliminary experiment in Sec. 6. Finally, we conclude this paper and discuss the future works in Sec. 7.

2. Related Works

There are the a number of difficulties in constructing an appropriate emotion corpus. Emotion is inherently a personal human experience and thus likely to be kept private. Furthermore, emotion is not something that can be replicated easily. This makes the collection of emotional data for research purposes a sensitive matter, often raising moral and ethical issues (El Ayadi et al., 2011). In addressing these difficulties, several methods and approaches have been inspected in previous studies; e.g. portrayal, simulation, and induction.

Acted emotion corpora contain recordings of actors portraying an assortment of emotion occurrences. In this set up, a number of actors are given a monologue script containing many different emotion expressions. They then read the lines and portray the emotion accordingly to the camera. The Geneva Multimodal Emotion Portrayals (GEMEP) corpus is one of the most widely used acted emotion corpus (Bänziger et al., 2006). Even though portrayals do provide emotionally rich data, the usage of acted affects as a base of study or experiment raises a number of concerns. These are due to the characteristics of acted emotions; it is stereotypical, lacking context, and limited to general and basic emotions.

To rid the disadvantages of acted emotion corpora, researchers tried to collect data from natural conversation. Two of the non-acted emotion corpora are the HUMAINE Database, a multimodal corpus consisting of natural and induced data showing emotion in a range of contexts (Douglas-Cowie et al., 2007), and the Vera am Mittag corpus, collected from German television broadcast (Grimm et al., 2008). The usage of interview recordings as emotional data is beneficial as they contain natural, more realistic emotion occurrences. However, they are often too subtle and contextualized due to the set up in which they are collected. Futhermore, even variety of emotions across the subjects can not be guaranteed.

On the other hand, the SEMAINE Database consists of natural dialogue between user and operator simulating a Sensitive Artificial Listener (SAL) with different traits (McKeown et al., 2012). 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 yet still allow for some control, but the set-up sometimes gives unpredictable result as the participants are fully aware of the intention of the agents.

Despite these new approaches, acted emotion remains one of the most promising bases of emotion studies due to its potential. Controls imposed on the corpus allows for comparison and analysis across different emotions and speakers. The prominent emotion content could be of advantage for preliminary study. Furthermore, a previous study had argued that careful design of research could balance the flaw of emotion portrayals (Bänziger and Scherer, 2007).

In this paper, we collect data from portrayal of various emotion occurrences by 14 native Japanese speakers. The novelty of our approach lies in the design of the portrayed emotion. In addition to the classic monologue reading, we include a number of situated dialogues. With the monologue, we gather prominent, simple, and isolated emotional speech, useful for preliminary multimodal emotion research. On the other hand, the dialogues are to provide less stereotypical, more subtle, and contextualized emotion, counteracting the shortcomings of typical emotion portrayal. Details of the corpus design and procedure of the construction will be described in the following sections.

3. Emotion Definition

Defining and structuring emotion is essential in observing and analyzing its occurrence. We define the emotion labels based on the circumplex model of affect (Russell, 1980). This model has been highly adopted and utilized in various emotion research. 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 five common emotion terms: *happiness*, *anger*, *sadness*, *contentment*, and *neutral*. In correspondence to the valence-arousal dimensions, happiness is positive-active, anger is negative-active, sadness is negative-passive, and contentment is positive passive. Neutral is associated with values of valence and arousal that are close to zero. Fig. 1 illustrates these emotional dimensions. This serves as the scope of emotion in this paper.

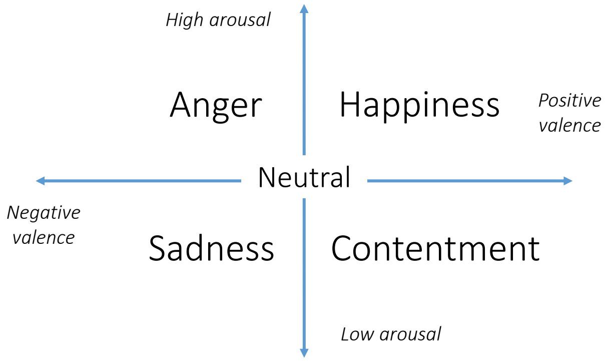

Figure 1: Emotion classes and dimensions

4. Data Collection

In this section, we describe the approach of the emotional data collection. We collected the data from scripted monologue and dialogue in Japanese. The script is tailored to evenly include the emotion classes described in Sec. 3. We also describe the recording set up of the data collection.

4.1. Script

One of the concerns raised on acted or portrayed emotion database is that the emotions are decontextualized. To address this, we divided the script into two parts: 1) monologue, and 2) dialogue.

Prior to recording, a script containing the emotional utterances is prepared. Each utterance is assigned one of the five common emotion terms (i.e. happy, angry, sad, content, neutral) as emotion label. We designed the script such that the defined emotions are evenly distributed in the corpus. Overall, the script contains 37 happy, 23 angry, 34 sad, 35 content, and 38 neutral utterances.

The monologue part contains isolated utterances, consists of 4 sentences per emotion class. This part is designed to give clean and prominent emotion occurrence that would be beneficial in preliminary emotion research experiments. We handcraft the sentences carefully, making sure to avoid emotion ambiguity in the semantic content, i.e. content that might suit more than one emotion. Table 1 contains a sample utterance for each emotion class from the monologue script.

Transcript	Emotion
Wow, that's great!	Happy
This is unfair!	Angry
My dog died last week.	Sad
It's nice to meet you.	Content
I was born on July 26.	Neutral

Table 1: A translated sample from the monologue script

On the other hand, the dialogue part contains short scenarios with different contexts. We adapt a number of examples of daily conversation initially designed for Japanese language learning purposes. These conversations are ideal as they are written to show common daily interaction, depicting relatable situation and emotion. As a result, the occurring emotion is contextualized and less stereotypical. A few adjustments are made on the collected conversations to ensure even incorporation all emotion classes. To name a few, the dialogue scenarios include a neighbor complaining about noise to another neighbor and a customer receiving a discount while shopping. Table 2 demonstrates one of the scenarios.

Speaker	Transcript	Emotion
A	*Excuse me, do you live in room 202?*	Angry
B	*Yes, I do.*	Neutral
A	*You know, I live right under your apartment. Could you stop using the washing machine at night? The noise keeps me up.*	Angry
B	*I'm sorry. I always get home late.*	Sad
A	*I understand that, but we're old and we want to go to sleep around 10.*	Angry
B	*I'm really sorry. I'll try to do my laundry in the morning whenever I can.*	Sad
A	*I didn't mean that. If you could do it before 10, doing it at night would be fine.*	Angry

Table 2: A translated sample from the dialogue script

4.2. Recording

We select 14 male Japanese native speakers to read and portray the script. This makes 7 pairs of speakers for the dialogue part. Recording was performed using two Kinect cameras and the built in 4-channel microphones. Kinect is chosen for recording for several reasons: (1) the built in microphone allows for simpler audio-video recording set up while still maintaining the quality of the result, and (2) it captures additional data that can be included in future corpus expansion, e.g. skeleton data for gesture and body language.

One session is carried per pair of speakers. The speakers face each other with a table between them. Two Kinects are installed on the table at a height that does not obstruct the speaker's view, each facing a different speaker. A foam separator is placed between the Kinects to ensure clean speech recording for each speaker. To allow natural reading for the camera, the script is displayed line by line behind the other speaker. The room set up is illustrated in Fig. 2.

Before starting the session, the speakers are briefed on the procedure and objective of the recording. After briefing, they are shown a clip of an example recording alongside a synchronized display of the script. This is intended to familiarize the speakers with the script and the expected emotion portrayal. After the speaker confirms, the recording is started.

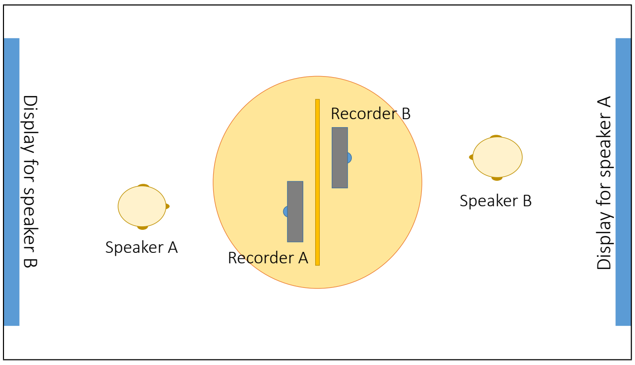

Figure 2: Recording set up

Fig. 3 shows a sample of sequence of face frames over a single utterance collected from the recording.

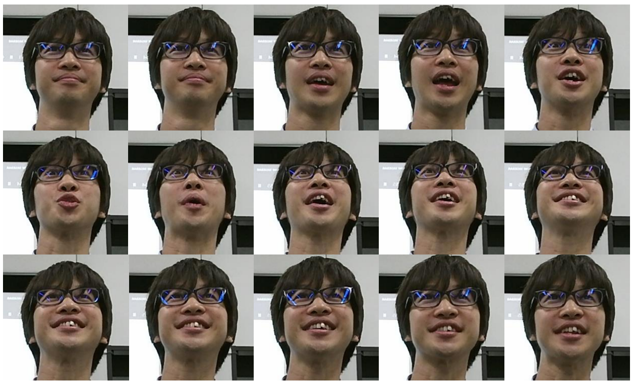

Figure 3: Face frames over a happy utterance

5. Annotation

In this section, we define and describe the annotation labels. We also explain the annotation procedure, where we impose thorough quality control to ensure the consistency of the results.

5.1. Emotion Labels

As previously mentioned in Sec. 4.1., the emotion class for each utterance is actually pre-determined in the script, eliminating the need for further emotion annotation. However, it is observed upon recording that the speakers portrayed these emotions differently. An angry utterance by one speaker may be more intense that that of other speaker, a sad utterance by one speaker may sound more depressed than that of other speaker, and so on. To capture these differences within an emotion class, we designed two sets of emotion labels according to the valence-arousal visualized in Fig. 1.

The emotion label sets are given in Table 3. We considered a set of *emotion dimension* labels in addition to *emotion class*. *Emotion dimension* set consists of the level of arousal and valence. The value of each dimension can be as low as -3 and as high as 3. For each emotion class, the value for the dimensions are bounded according to Fig. 1, i.e. for anger, valence ranges from -1 to -3, and activation ranges from 1 to 3, and so on. In other words, the emotion

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

Table 3: *Emotion label sets*

dimension labels serve as a more fine-grained information of the emotion class labels.

5.2. 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 a native Japanese speaker to annotate the full corpus.

Fig. 4 gives an overview of the annotation procedure. Before annotating the corpus, the annotator is 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 as well as a number of examples.

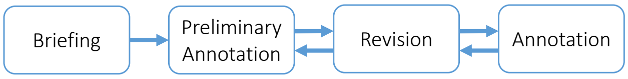

Figure 4: Overview of annotation procedure

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

We manually screen the preliminary annotation result and give feedback to the annotator accordingly. The annotator is asked to revise inconsistencies with the guidelines if there are any. For example, when valence and arousal value of an emotion class is not according to the definition. This revision is important in ensuring a consistent emotion description in the annotation. We perform the same screen-and-revise process on the full corpus annotation to achieve a tenable result.

6. Emotion Recognition

We perform preliminary experiment of speech-based emotion recognition with the collected data. The data is partitioned with 85:15 ratio for training set and test set. We performed three different recognition schemes: (1) emotion class recognition, (2) valence level recognition, and (3) arousal level recognition.

We extracted baseline acoustic feature set from INTERSPEECH 2009 emotion recognition challenge (Schuller et al., 2009) using the openSMILE toolkit (Eyben et al., 2010). This feature set is described in Table 4. In total, 384 features are extracted for each utterance as classification features. We then test 3 different algorithm for the recognition: Support Vector Machine (SVM) by means of libSVM (Chang and Lin, 2011), log regression, and neural network (NN). Fig. 5 visualizes the result.

LLD (16 · 2)	Functionals (12)
(△) ZCR	mean
(△) RMS Energy	standard deviation
(△) F0	kurtosis, skewness
(△) HNR	extremes: value, rel. position, range
(△) MFCC 1-12	linear regression: offset, slope, MSE

Table 4: Baseline feature of INTERSPEECH 2009 emotion challenge

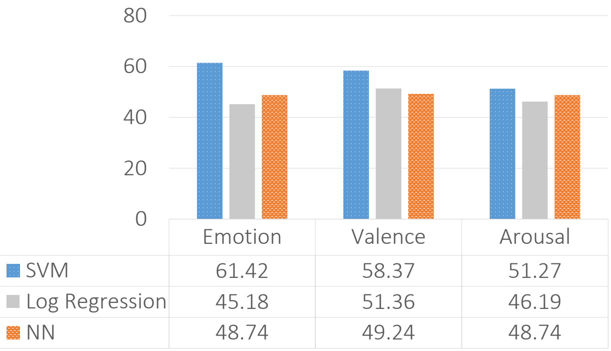

Figure 5: Performance of automatic recognition (in %)

In this experiment, SVM outperforms log regression and NN on all recognition schemes. The same SVM procedure was previously performed on (Lubis et al., 2014) on the SEMAINE database, where emotion recognition accuracy of 52.08% was achieved for four emotion classes. Given the same technique of recognition and the significantly fewer data compared to SEMAINE, this experiment could give an insight to the quality of emotion occurrences contained in the corpus.

7. Conclusion and future works

We presented an audio-visual emotion corpus in Japanese. We collected approximately 100 minutes of material from emotion portrayals of 14 Japanese native speakers. The recording was performed in monologue and dialogue format to provide both simple emotion occurrences and contextualized ones. We carefully annotated the data using two sets of labels to preserve the details of differences of the emotion occurrences. In this paper, we also presented the result of the preliminary experiments performed using the corpus.

In the future, we look forward to increase the size of the corpus by incorporating more speakers and more scenarios. Addition of a role-play dialogue simulation could provide a more natural yet still controlled material. The different type of emotion occurrences contained in the corpus can allow for a meaningful analysis. We are hoping to perform a multimodal emotion recognition on the corpus, as well as improving the emotion recognition accuracy with the additional data.

8. Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Number 26870371.

9. Bibliographical References

Bänziger, T. and Scherer, K. R. (2007). Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In *Affective computing and intelligent interaction*, pages 476–487. Springer.

Bulut, M., Narayanan, S. S., and Syrdal, A. K. (2002). Expressive speech synthesis using a concatenative synthesizer. In *INTERSPEECH*.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Iida, A., Campbell, N., Higuchi, F., and Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication*, 40(1):161–187.

Kitanaka, J. (2011). *Depression in Japan: Psychiatric cures for a society in distress*. Princeton University Press.

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

Petrantonakis, P. C. and Leontios, J. (2014). EEG-based emotion recognition using advanced signal processing techniques. *Emotion Recognition: A Pattern Analysis Approach*, pages 269–293.

Reeves, B. and Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Schuller, B., Steidl, S., and Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315.

Wang, W., Athanasopoulos, G., Patsis, G., Enescu, V., and Sahli, H. (2014). Real-time emotion recognition from natural bodily expressions in child-robot interaction. In *Computer Vision-ECCV 2014 Workshops*, pages 424–435. Springer.

10. Language Resource References

Arimoto, Y and Kawatsu, H and Ohno, S and Iida, H. (2008). *Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems*.

Bänziger, Tanja and Pirker, Hannes and Scherer, K. (2006). *GEMEP-Geneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions*.

Douglas-Cowie, Ellen and Cowie, Roddy and Sneddon, Ian and Cox, Cate and Lowry, Orla and Mcrorie, Margaret and Martin, Jean-Claude and Devillers, Laurence and Abrilian, Sarkis and Batliner, Anton and others. (2007). *The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data*.

Grimm, Michael and Kroschel, Kristian and Narayan, Shrikanth. (2008). *The Vera am Mittag German Audio-Visual Emotional Speech Database*.

Lyons, Michael J and Akamatsu, Shigeru and Kamachi, Miyuki and Gyoba, Jiro and Budynek, Julien. (1998). *The Japanese female facial expression (JAFFE) database*.

McKeown, Gary and Valstar, Michel and Cowie, Roddy and Pantic, Maja and Schroder, Marc. (2012). *The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent*.

Zhang, Hao and Lopez, Guillaume and Shuzo, Masaki and Omiya, Yasuhiro and Mistuyoshi, Shunji and Warisawa, Shin'ichi and Yamada, Ichiro. (2014). *A Database of Japanese Emotional Signals Elicited by Real Experiences*.