一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2016-51 (2016-12)

感情的独話・対話を用いた日本語マルチモーダルコーパスの構築

ヌルルルービス[†]　ランディゴメス[††]　サクリアニサクティ[†]　中村　圭佑[††]　吉野幸一郎[†]
中村　　哲[†]　中臺　一博[††]

[†] 奈良先端科学技術大学院大学
[††] ホンダ・リサーチ・インスティチュート・ジャパン

あらまし　ヒューマンコンピュータインタラクションにおいて感情を用いるため、様々な種類のモダリティにおけるラベルを付与したデータが必要となる。しかし日本語における感情コーパスでは、多くのコーパスが 1 種類または 2 種類のモダリティに限られている。そこで本研究では、感情の生起を様々な側面から観測するため、音声および映像からなる日本語感情コーパスを作成した。本コーパスは 14 人の日本語母語話者を対象に収録した。また、収録したコーパスに対して音声のモダリティを用いた感情認識タスクを行い、5 感情の分類において 61.42%の精度を実現した。

キーワード　マルチーモーダル、感情、コーパス、日本語

Constructing a Japanese Multimodal Corpus From Emotional Monologues and Dialogues

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.

Abstract　To fully incorporate emotion into human-computer interaction, rich sets of labeled emotional data is prerequisite. However, in Japanese, the majority of the existing emotion database is still limited to unimodal and bimodal corpora. To allow more complete observation of emotion occurrences, we construct the first audio-visual emotion corpora in Japanese, collected from 14 native speakers. Preliminary speech emotion recognition experiments on the corpus and achieved an accuracy of 61.42% for five classes of emotion.

Key words　multimodal, emotion, corpus, Japanese

1. Introduction

Emotion sensitive systems has great potential in enhancing human-computer interaction. 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 [1], [2]. In all of these efforts, emotion corpus is a prerequisite.

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 [3], and the aging population are among many conditions where assistive technology is in serious need. An emotion sensitive system supporting intensive care and treatment will be of valuable aid in addressing these issues.

However, in Japanese, the majority of existing emotion corpora are still limited to textual, speech [4], and physiological signals [5]. These resources are still missing one of the most important outlet of emotion expression: visual. In this paper, we present an audio-visual emotion corpus in Japanese. We collected portrayal of various emotion occurrences from 14 native speakers. The corpus contains approximately 100 minutes of annotated and transcribed material.

2. Emotion Definition

We define the emotion labels based on the circumplex model of affect [6]. 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

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.

emotion terms: happiness, anger, sadness, contentment, and neutral. This serves as the scope of emotion in this paper.

3. Data Collection

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. 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.

The monologue part contains isolated utterances, consists of 4 sentences per emotion labels. The monologue part is designed to give simple and basic emotion occurrence. On the other hand, the dialogue part contains short scenarios with different contexts, e.g. a neighbor complaining about noise to another neighbor. These scenarios are intended to give more contextualized emotion occurrences. As a result, the occurring emotion is more complex and less stereotypical. Overall, the script contains 37 happy, 23 angry, 34 sad, 35 content, and 38 neutral utterances.

We select 14 male Japanese native speakers to read and portray the script, making 7 dialogue pairs. One session is carried per pair of speakers. The speakers stand facing each other with two Kinects are set up in between them. The script is displayed line by line behind the other speaker to allow a natural reading for the camera.

4. Annotation

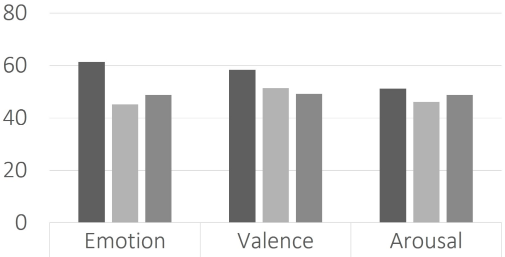
To capture the differences within an emotion class, we consider a set of *emotion dimension* labels in addition to *emotion class*. *Emotion dimension* set consists of the level of arousal and valence. The value of each dimension can be as low as -3 and as high as 3. For each emotion class, the value for the dimensions are bounded according to the emotion definition, e.g. for anger, valence ranges from -1 to -3, and activation ranges from 1 to 3. In other words, the emotion dimension labels serve as a more fine-grained information of the emotion class labels.

We carefully select one native Japanese speaker to annotate the full corpus. Before annotating the corpus, the annotator is briefed and given a document of guidelines explaining the task and its goal. After briefing, we ask the annotator to do preliminary annotation by working on a small subset of the corpus 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. The annotator is asked to revise inconsistencies with the guidelines if there are any. 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.

5. 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 using the openSMILE toolkit. 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), log regression, and neural network (NN). Fig. 1 visualizes the result.



	Emotion	Valence	Arousal
■ SVM	61.42	58.37	51.27
■ Log Regression	45.18	51.36	46.19
■ NN	48.74	49.24	48.74

Figure 1 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 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.

6. Conclusion and future works

We presented an audio-visual emotion corpus in Japanese. 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. Preliminary speech emotion recognition experiments on the corpus and achieved an accuracy of 61.42% for five classes of emotion. 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.

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," Computer Vision-ECCV 2014 WorkshopsSpringer, pp.424–435 2014.

[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] J. Kitanaka, Depression in Japan: Psychiatric cures for a society in distress, Princeton University Press, 2011.

[4] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, "Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems," 2008.

[5] H. Zhang, G. Lopez, M. Shuzo, Y. Omiya, S. Mistuyoshi, Shin’ichiWarisawa, I. Yamada, "A database of Japanese emotional signals elicited by real experiences," 2014.

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