INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Emotion category mapping to emotional space by cross-corpus emotion labeling

Yoshiko Arimoto[1,2], *Hiroki Mori*[3]

[1]Brain Science Institute, RIKEN, Japan
[2]Faculty of Science and Engineering, Teikyo University, Japan
[3]Graduate School of Engineering, Utsunomiya University, Japan

ar@brain.riken.jp, hiroki@speech-lab.org

Abstract

The psychological classification of emotion has two main approaches. One is emotion category, in which emotions are classified into discrete and fundamental groups; the other is emotion dimension, in which emotions are characterized by multiple continuous scales. The cognitive classification of emotion by humans perceived from speech is not sufficiently established. Although there have been several studies on such classification, they did not discuss it deeply. Moreover, the relationship between emotion category and emotion dimension perceived from speech is not well studied. Aiming to establish common emotion labels for emotional speech, this study elucidated the relationship between the emotion category and the emotion dimension perceived by speech by conducting an experiment of cross-corpus emotion labeling with two different Japanese dialogue corpora (Online Gaming Voice Chat Corpus with Emotional Label (OGVC) and Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB)). A likelihood ratio test was conducted to assess the independency of one emotion category from the others in three-dimensional emotional space. This experiment revealed that many emotion categories exhibited independency from the other emotion categories. Only the neutral states did not exhibit independency from the three emotions of sadness, disgust, and surprise.

Index Terms: cross corpus emotion labelling, emotional speech, emotion perception

1. Introduction

In research on emotion recognition by speech, the use of multiple large-scale speech corpora with a common emotion label is needed to test its effectiveness. However, two different corpora cannot be used together because emotion labels are exclusively labeled for either of the corpora based on their own criteria; there is no common shared labeling for both of them. A more crucial problem is that different emotion labeling schemes are adopted by different speech corpora. There are two primary emotion labels based on one of two different psychological emotion theories. One is emotion category theory, which claims that emotion is a discrete internal state such as joy or sadness, such as Ekman's big six emotions [1] or Plutchik's primary eight emotions [2]. The other is emotion dimension theory, which claims that emotion is a continuous internal state of several dimensions such as pleasant–unpleasant or aroused–sleepy, such as Russell's circumplex model [3] . When one emotional speech corpus is labeled with a different emotion label based on different labeling schemes, it is not possible to use both corpora in the same study. Even if one corpus is labeled with the same emotion labels as the other, the same emotion labels are not considered to be equivalent between the two corpora.

Although the same emotion labels cannot be equivalent between multiple corpora, several researches have examined emotion recognition and emotional speech synthesis with multiple corpora [4, 5, 6, 7, 8, 9]. Schuller et al. used eight emotional speech corpora for their research [6, 7, 8, 9]. The emotion labels for each of the eight corpora varied: one used four emotion categories, another used two emotion dimensions, another used two emotion categories, and so on. Those various emotion labels were classified by researchers into one of four quadrants of orthogonal two-dimensional space (pleasant–unpleasant and aroused–sleepy) to obtain ground-truth labels for speeches. This approach of adopting multiple corpora does not guarantee the equivalency of emotion labels among corpora. Zong et al. used four corpora for their emotion recognition research by selecting speech that was labeled with the same emotions across the four corpora. However, as mentioned above, this method also does not guarantee the equivalency of emotion labels across corpora and allowed the use of only a limited number of utterances in the four corpora. A standardization of common emotion labels across emotional speech corpora is required.

With the aim of standardizing the same emotion labels across speech corpora, we first elucidate the relationship between well-known emotion labels, i. e., emotion category and emotion dimension. Using two publicly available Japanese dialog speech corpora with emotion labels, we conducted cross-corpus emotion labeling to label utterances of the two corpora with both the emotion category label and the emotion dimension label. Then, a likelihood ratio test was conducted to assess the independency of one emotion category from the others in three-dimensional emotional space.

2. Speech material

This section introduces the two publicly available Japanese dialog speech corpora used for our research. One is the Online Gaming Voice Chat Corpus with Emotional Label (OGVC) [10] and the other is the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB) [11].

2.1. Online Gaming Voice Chat Corpus with Emotional Label (OGVC)

OGVC was developed specifically for research on emotion recognition and emotional speech synthesis. It includes two types of emotional speech material, spontaneous dialog speech and acted speech which has the same linguistic content as the

http://dx.doi.org/10.21437/Interspeech.2017-994

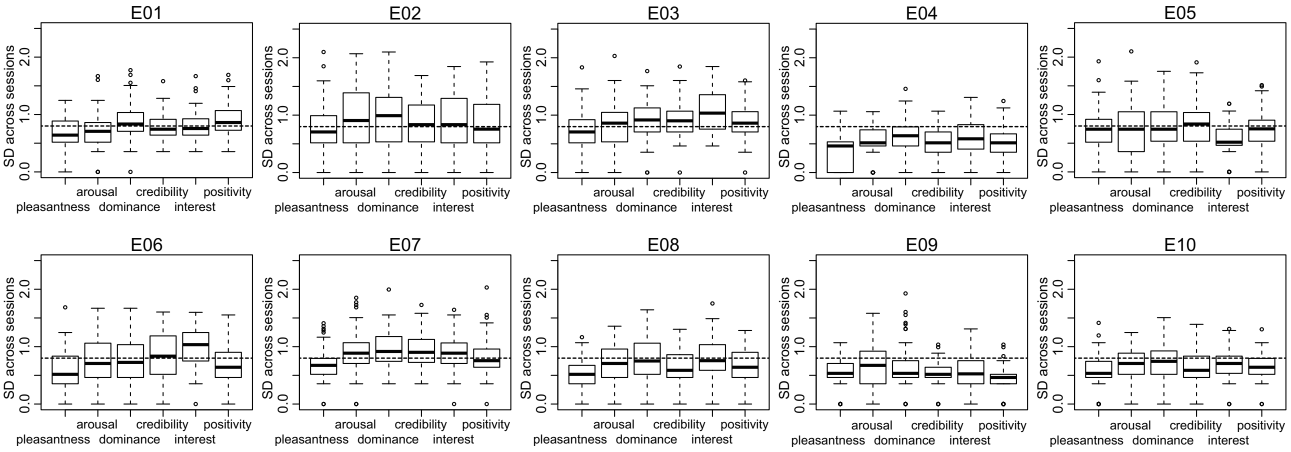

Figure 1: *Intra-labeler SD of six emotion dimension evaluation.*

spontaneous dialog speech. In this study, we used only the spontaneous dialog speech, which was recorded while two or three speakers were chatting over the Internet while participating in a Massively Multi-player Online Role-Playing Game (MMORPG). The spontaneous dialog speech corpora of OGVC includes 9,114 utterances of 13 Japanese speakers (4 females and 9 males). The utterance of this corpus was defined based on 400 ms inter-pausal units (IPUs). 6,578 out of all the utterances have one of 10 emotion category labels, i. e., joy (JOY), acceptance (ACC), fear (FEA), surprise (SUR), sadness (SAD), disgust (DIS), anger (ANG), anticipation (ANT), neutral (NEU), and others (OTH). Eight of the 10 emotion category labels were based on Plutchik's primary emotions [2]. The 6,578 emotion-labeled utterances were used for the experiment.

2.2. Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB)

UUDB is a collection of natural, spontaneous dialogs of Japanese college students. The participants engaged in the "four-frame cartoon sorting" task, where four cards each containing one frame extracted from a cartoon are shuffled, and each participant has two cards out of the four, and is asked to estimate the original order without looking at the remaining cards. The current release of the UUDB includes dialogs of seven pairs of college students (12 females, 2 males), composed of 4,840 utterances. The utterance was defined as a speech continuum bounded by either silence ($>$ 400 ms) or slash unit boundaries. For all utterances, perceived emotional states of speakers are provided. The emotional states were annotated with the following six abstract dimensions:

- pleasant–unpleasant
- aroused–sleepy
- dominant–submissive
- credible–doubtful
- interested–indifferent
- positive–negative

The emotional state for each utterance was evaluated on a seven-point scale. In evaluating the pleasant–unpleasant scale, for example, 1 corresponds to extremely unpleasant, 4 to neutral, and 7 to extremely pleasant. All 4,840 utterances were used for the experiment.

3. Labeler screening

This section introduces a method of labeler screening for cross-corpus emotion labeling. As described in Section 2, OGVC and UUDB use different emotion labels, and so the two corpora cannot be used together for any research. The emotion labels included in the delivered corpora were discarded, and the emotion category and emotion dimension were originally labeled for all the utterances of both OGVC and UUDB in our experiment. To obtain original, common emotion labels across two corpora, qualified labelers should be screened out.

3.1. Procedure

Ten undergraduate or graduate students (6 females and 4 males, mean age 21.8 (SD 0.84)) participated in the screening test.

The emotion labeling framework of both emotion category and emotion dimension were the same as those in the two distributed corpora. In emotion category labeling, the labelers had to choose one of 10 categories (JOY, ACC, FEA, SUR, SAD, DIS, ANG, ANT, NEU, and OTH) for each utterance. The grand truth label for each utterance was determined by the majority vote of the labelers. In emotion dimension labeling, the labelers had to rate six emotion dimensions on a seven-point scale for each utterance. The grand truth label of emotional dimension for each utterance was defined as the mean score of the labelers. Each labeler conducted both emotion category and emotion dimension labeling tasks. The emotion dimension labeling task preceded the emotion category labeling task.

One hundred eight utterances from two corpora were selected for the screening test: 54 were obtained from OGVC (9 emotions × 3 emotional intensity levels (weak, middle, and strong) × 2 utterances) and the remaining 54 from UUDB (3 emotional intensity levels (weak, middle, and strong) × 18 utterances). The emotion labeling for the labeler screening consisted of 8 repeated sessions × one block of 108 utterances × 2 types of labeling (category and dimension).

3.2. Analysis

To check each labeler's consistency of evaluation, intra-labeler standard deviation of each six-emotion dimension labeling was calculated across eight repeated sessions.

To choose the labelers who had the same emotional sensitivity as each other, the agreement between each labeler's evaluation and the correct label was calculated for emotion category

Table 1: *The agreement of emotion category and the correlation coefficient of emotion dimension between each labeler and mean label across labelers*

Evaluator	Agreement	r
E01	0.42	0.82
E02	0.52	0.89
E03	0.44	0.85
E04	0.60	0.90
E05	0.50	0.86
E06	0.53	0.83
E07	0.51	0.81
E08	0.42	0.83
E09	0.36	0.59
E10	0.48	0.87

labeling. The correct label for each utterance was defined by majority voting for this screening test. Likewise, the correlation coefficients between each labeler's evaluation and the correct label were calculated for emotion dimension labeling. The correct label for each utterance was defined by mean evaluation across 10 labelers for this screening test.

3.3. Results

Figure 1 shows the mean SDs of the six emotion dimension evaluation for each labeler. A horizontal dashed line at 0.8 on the y-axis for each panel is a criterion for the labeler's consistency based on [11]. The labelers whose SDs for all emotion dimensions were less than the criteria were E04, E08, E09, and E10. Table 1 shows the agreement of emotion category and the correlation coefficient of emotion dimension between each labeler and mean label. The labelers who exhibited the highest agreement on emotion category labeling were E04 (0.60), E06 (0.53), E02 (0.52), and E07 (0.51). The labelers who exhibited the highest correlation coefficient on emotion dimension labeling were E04 (0.90), E02 (0.89), E10 (0.87), E05 (0.86).

3.4. Discussion

To obtain labelers whose evaluations were consistent, all the labelers were screened based on the criterion of labeler's consistency (all SDs for the six emotion dimensions were less than 0.8). As a result, E04, E08, E09, and E10 were selected as candidate labelers for cross-corpus emotion labeling. Among the four labelers, E04 exhibited the highest agreement on emotion labeling and the highest correlation coefficient on emotion dimension labeling. Thus, E04 was selected as the primary labeler for the cross-corpus emotion labeling. Among the other three labelers, E09 exhibited the lowest agreement and the lowest correlation coefficient, and so was screened out for the cross-corpus emotion labeling. As a result, E04, E08, and E10 were selected as the labelers for the following experiment.

4. Cross-corpus emotion labeling and emotion category mapping to emotional space

This section introduces the procedure of cross-corpus emotion labeling. As described in Section 3, the emotion labels included in the delivered corpus were discarded, and the emotion category and emotion dimension were originally labeled for the utterances of OGVC and UUDB in our experiment.

Table 2: *The number of utterance each emotion category*

Emotion	OGVC	UUDB	Total
JOY	438	259	697
ACC	623	1030	1653
FEA	282	94	376
SUR	313	120	433
SAD	488	331	819
DIS	970	406	1376
ANG	128	39	167
ANT	186	59	245
NEU	18	13	31
Total	3446	2351	5797

4.1. Procedure

The three qualified labelers, selected by the labeler screening, conducted the cross-corpus emotion labeling. The mean age of the three labelers was 22 years old (SD 0.82). The same emotion labels as used in the labeler screening test were used for the cross-corpus emotion labeling. Each labeler evaluated 11,418 utterances from both OGVC and UUDB (6,578 from OGVC and 4,840 from UUDB). The 11,418 utterances were randomly separated into blocks. The cross-corpus emotion labeling consisted of 104 blocks of 11,418 utterances × 2 types of labeling (category and dimension).

4.2. Analysis

For emotion category labeling, a majority vote among the three labelers was regarded as the correct emotion label for each utterance. For emotion dimension labeling, the mean rating of the three labelers was regarded as the correct intensity of each emotion dimension.

To assess the independency of one emotion category from the others on n-dimensional emotional space, an equivalence test for two n-dimensional Gaussian mixture models (GMMs) was conducted. For each pair of two emotion categories E_1 and E_2, the n-dimensional variables X_1 and X_2 belonging to each category were assumed to be generated from their corresponding GMM. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the sub-dataset belonging to E_1 and E_2, and N_1 and N_2 be the data size, respectively. The null hypothesis (H_0) and the alternative hypothesis (H_1) are as follows:

H_0: All instances of X_1 are generated from a GMM M_1 and all instances of X_2 are generated from a GMM M_2, which is identical to M_1.

H_1: All instances of X_1 are generated from a GMM M_1 and all instances of X_2 are generated from a GMM M_2, which differs from M_1.

The null hypothesis can be tested using the parametric bootstrap likelihood ratio test, where the distribution of the difference of the deviances (-2 times the log likelihood ratio) between the null model (M_1 and M_2 are trained, as an identical model, from random samples of size $N_1 + N_2$) and the alternative model (M_1 and M_2 are trained separately from random samples of size N_1 and random samples of size N_2) is estimated by random sampling under H_0. If the difference of the deviances between the null model (trained from $\mathbf{x}_1 + \mathbf{x}_2$) and the alternative model (trained separately from $\mathbf{x}_1$ and $\mathbf{x}_2$) falls into the critical region ($\alpha = 5\%$), then the null hypothesis is rejected and the two emotion categories are independently distributed on the n-dimensional emotional space. The likelihood ratio tests were conducted for all combinations of nine emotion categories.

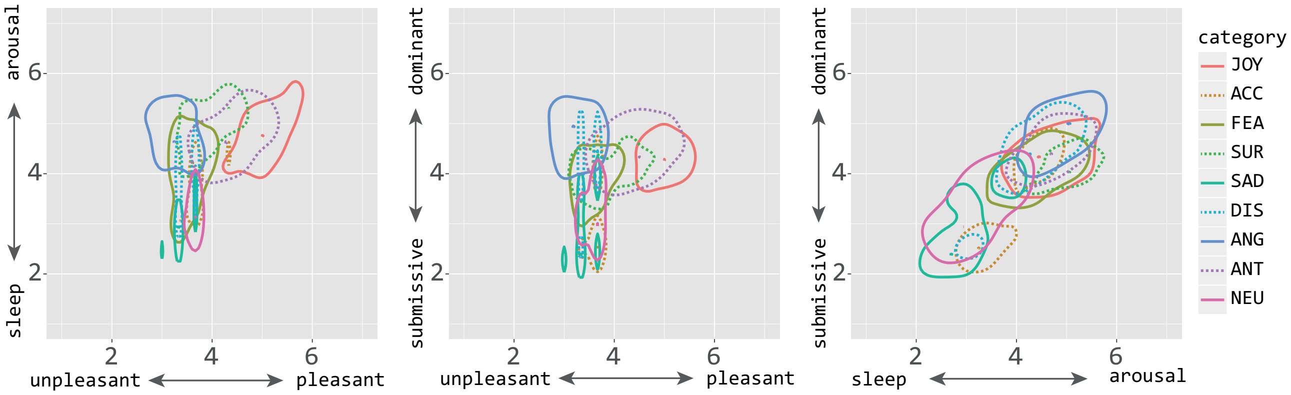

Figure 2: *Distribution of emotion categories on two-dimensional emotional space*

Table 3: *Difference of deviance between emotion categories mapping on three-dimensional emotional space*

	ACC	FEA	SUR	SAD	DIS	ANG	ANT	NEU
JOY	1187.3*	762.7*	680.1*	1406.7*	1660.8*	679.6*	178.0*	159.1*
ACC		501.8*	585.3*	1248.5*	1169.5*	765.3*	368.4*	367.2*
FEA			98.3*	342.9*	123.7*	215.2*	268.4*	41.7*
SUR				802.0*	482.7*	287.7*	253.2*	31.0
SAD					534.4*	603.8*	678.8*	38.2
DIS						108.6*	463.0*	11.6
ANG							361.6*	101.6*
ANT								99.8*

4.3. Results

Table 2 shows the number of utterances for each emotional category as a result of emotion category labeling. The total number of utterances for which two out of the three labelers agreed with one emotion label was 5,797 utterances (3,446 for OGVC and 2,351 for UUDB). It is 51% of the total utterances submitted to the cross-corpus labeling (52% of OGVC and 49% of UUDB). ACC, DIS, JOY and SAD were the major emotions by number of utterances in descending order. Hereafter, the 5,797 utterances were submitted for analysis of the emotion category mapping to emotional space.

Figure 2 shows the distribution of emotion category to the two-dimensional emotional spaces of pleasantness vs. arousal, pleasantness vs. dominance, and arousal vs. dominance. Table 3 shows the difference of deviances between emotion category mapping on three-dimensional emotional space. The asterisks in Table 3 indicate the combinations of emotion categories for which the hypothesis H_0 was rejected and the hypothesis H_1 was accepted ($p < 0.05$). The tests for many combinations of emotion categories rejected H_0; only three tests for NEU did not reject H_0 when testing with SUR, SAD, and DIS.

4.4. Discussion

In the pleasantness vs. arousal space in the left panel of Fig. 2, JOY (solid red line in Fig. 2) is placed on the upper right quadrant of high arousal and high pleasantness, SUR (dashed green line) on high arousal, SAD (solid green line) on low arousal, and ANG (solid blue line) on the upper left of high arousal and low pleasantness. These distributions are similar to Russell's circumplex model [3]. The result also shows that NEU (solid purple line) was placed around 4 on the pleasantness axis, but was placed from 2 to 4 on both the arousal and dominance axes. NEU is generally considered to be an emotionally neutral state placed at a score of 4 on any emotion dimension. However, our result implies that neutral utterance is neutral on the pleasantness dimension but is not necessarily neutral on the other dimensions.

The result of the likelihood ratio test on the distribution of each emotion category on three-dimensional emotional space suggested that all the combinations of emotion categories except NEU–SUR, NEU–SAD, and NEU–DIS exhibited significant differences between each other ($p < 0.05$). In other words, all the emotion categories except NEU are independent of each other. This result suggests that the information of eight emotion categories (JOY, ACC, FEA, SUR, SAD, DIS, ANG, and ANT) was not lost even in emotion dimension perception.

5. Conclusions

Aiming to standardize emotion labels across speech corpora, we first studied the relationship between emotion category and emotion dimension. Using two Japanese dialog speech corpora with emotion labels, cross-corpus emotion labeling was conducted to label utterances of the two corpora with both an emotion category label and an emotion dimension label. Then, a likelihood ratio test was conducted to assess the independency of one emotion category from the others on three-dimensional emotional space.

The test revealed that all the combinations of emotion categories except neutral–surprise, neutral–sadness, and neutral–disgust exhibited significant differences between each other. All the emotion categories except neutral were independent of each other in the dimensional emotional space.

The result suggested the surprising fact that the information of eight emotion category including joy, acceptance, fear, surprise, sadness, disgust, anger, and anticipation, was not lost even in emotion dimension perception. However, another future research with another language speech corpora may exhibit different result from this research, because emotion perception heavily depends on its language, culture or social norm. The universal standardization of emotion labeling will be accomplished after examining the linguistic difference, cultural difference, or social difference of standard emotion label.

Further research is planned to elucidate the influence of acoustic information on emotion category and emotion dimension perception.

6. Acknowledgements

This work was supported by TATEISHI Science and Technology Foundation Grant for Research (A) and JSPS KAKENHI Grant Number 26280100.

7. References

[1] P. Ekman and W. V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions*. New Jersey: Prentice Hall, 1975.

[2] R. Plutchik, *Emotions: A psychoevolutionary synthesis*. New York: Harper & Row, 1980.

[3] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[4] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-Corpus Speech Emotion Recognition Based on Domain-Adaptive Least-Squares Regression," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 585–589, may 2016.

[5] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, 2016.

[6] B. Schuller, Z. Zhang, F. Weninger, and F. Burkhardt, "Synthesized speech for model training in cross-corpus recognition of human emotion," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 313–323, 2012.

[7] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings*, pp. 523–528, 2011.

[8] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-Corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[9] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009*, pp. 552–557, 2009.

[10] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, "Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment," *Acoustical Science and Technology*, vol. 33, no. 6, pp. 359–369, 2012.

[11] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, aug 2011.