Challenges of building an authentic emotional speech corpus of spontaneous Japanese dialog

Challenges of Building an Authentic Emotional Speech Corpus of Spontaneous Japanese Dialog

Yoshiko Arimoto
Faculty of Science and Engineering, Teikyo University
1-1 Toyosato, Utsunomiya, Tochigi, Japan
ar@mac-lab.org

Abstract

This paper introduces the challenges involved in studying authentic emotional speech collected from spontaneous Japanese dialog. First, three key issues related to emotional speech corpora are presented: data type (acted or spontaneous), efficient collection of emotional speech, and appropriate emotion labeling. To address these issues, a data collection scheme was developed, and a labeling experiment was performed. First, a data collection scheme using an online game task was applied to efficiently collect speakers' authentic emotional expressions during their real-life conversations. Then, to elucidate appropriate emotion labels for emotional speech and to commonize the emotion labels among several corpora, the relationship between emotion categories and emotion dimensions, which are two major approaches to psychological emotional modeling, was demonstrated by conducting a cross-corpus emotion labeling experiment with two different Japanese dialogue corpora (the Online Gaming Voice Chat Corpus with Emotional Label (OGVC) and the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB)). Finally, the results are presented, and the advantages and disadvantages of these approaches are discussed.

Keywords: emotional speech corpus, Japanese dialog speech, data collection, emotion labeling

1. Introduction

Emotional speech has been studied to elucidate its acoustic profiles and for applications in automatic emotion recognition and emotional speech synthesis. Various emotional speech corpora have been used for such studies. Emotional speech corpora can be classified into two types based on how the speech is produced: acted emotional speech corpora and authentic emotional speech corpora. Many of the studies on emotional speech have used acted emotional speech to investigate the acoustical correlation with emotion (Williams, 1972; Itoh, 1986; Kitahara, 1988; Banse and Scherer, 1996; Engberg et al., 1997). Such acted speech consists of idealized speech samples generated to match someone's conception of what an emotion should be like (Cowie, 2009), with well-designed prosodic and acoustic expression recorded in the noiseless environment of a soundproof room.

The contrast to acted emotional speech is authentic emotional speech. For practical applications such as automatic emotion recognition research and emotional or expressive speech synthesis, speech corpora containing authentic emotional speech samples evoked during real-life conversation are indispensable because such applications are designed for a real-world environment, not a laboratory setting. Several research groups began to study spontaneous emotional speech in the late 1990s (Ang et al., 2002; Arimoto et al., 2007). In that research, several attempts were made to record the expression of authentic emotions during spontaneous dialogs: dialogs between the AutoTutor system and students (Litman and Forbesriley, 2006), dialogs between a robotic pet and a child (Batliner et al., 2011), and interviews in which the speaker's emotions were controlled by the experimenter (Douglas-Cowie, 2003). Devillers and Vidrascu investigated real conversations during telephone calls with a call center (Devillers et al., 2006). In addition, several studies on authentic emotional speech have been performed with spontaneous materials (Campbell, 2004; Arimoto et al., 2008; Mori et al., 2011). Zeng et al. (Zeng et al., 2009) and Cowie (Cowie, 2009) have presented detailed reviews of the history of emotional speech corpora and suggestions for constructing an emotional speech corpus.

However, some issues arise with regard to the use of authentic emotional speech samples collected from spontaneous dialog. One issue is the data type: acted speech or spontaneous speech. Cowie (Cowie, 2009) demonstrated an example of the implications of this issue by means of a meta-analysis of automatic emotion recognition. The recognition rate using authentic emotional speech is lower than that using acted emotional speech. This report suggested that authentic emotional speech acoustically differs from acted emotional speech. Jürgens et al. supported this suggestion by identifying acoustic differences between authentic emotional speech and acted speech (Jürgens et al., 2011). Moreover, a method trained on acted speech, with deliberately and exaggeratedly expressed emotion, failed to generalize to authentic speech with subtle and complex emotional expression (Batliner et al., 2003; Zeng et al., 2009). Another critical issue noted with respect to spontaneous materials is the quantity of authentic emotional speech collected during spontaneous dialog. Cowie observed that even a large speech corpus contains few emotional samples (Cowie, 2009). Campbell recorded telephone conversations and labeled each recorded utterance with an observed emotion (Campbell, 2004). Although real-life conversations were successfully recorded, little of the speech displayed strong emotional content. Ang et al. (Ang et al., 2002) also obtained little emotional speech, although approximately 22,000 utterances were collected from a pseudodialog. Those studies suggested that methods of evoking emotion are necessary to efficiently collect

authentic emotional speech from spontaneous dialog.

Another issue is emotion labeling for authentic emotional speech. In research on emotion recognition from speech, the use of multiple large-scale speech corpora with common emotion labels is needed to test the effectiveness of recognition. However, two different corpora typically cannot be used together because the emotion labels for each of the corpora are assigned 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 for different speech corpora. There are two primary types of emotion labels, each based on one of two different psychological emotion theories. One is emotion category theory, which claims that emotions are discrete internal states such as joy or sadness, such as Ekman's Big Six emotions (Ekman and Friesen, 1975) or Plutchik's eight primary emotions (Plutchik, 1980). The other is emotion dimension theory, which claims that emotion is a continuous internal state with several dimensions, such as pleasant–unpleasant and aroused–sleepy, as described by Russell's circumplex model (Russell, 1980), for example. When different emotional speech corpora are labeled with different emotion labels based on different labeling schemes, it is not possible to use both corpora in the same study. Even if two corpora are labeled with emotion labels of the same type, the emotion labels are not considered to be equivalent between the two corpora.

Although the emotion labels cannot be equivalent among multiple corpora, several researchers have examined emotion recognition and emotional speech synthesis with multiple corpora (Zong et al., 2016; Song et al., 2016; Schuller et al., 2012; Zhang et al., 2011; Schuller et al., 2010; Schuller et al., 2009). Schuller et al. used eight emotional speech corpora in their research (Schuller et al., 2012; Zhang et al., 2011; Schuller et al., 2010; Schuller et al., 2009). The emotion labels for each of the eight corpora varied: one used four emotion categories, another used two emotion dimensions, another used two different emotion categories, and so on. The various emotion labels were classified by the researchers into one of four quadrants of an orthogonal two-dimensional space (pleasant–unpleasant and aroused–sleepy) to obtain ground-truth labels for the speech samples. However, this approach to using multiple corpora does not guarantee the equivalency of the emotion labels among the corpora. Zong et al. used four corpora for emotion recognition research by selecting speech samples that were labeled with the same emotions across all four corpora. However, this method also does not guarantee the equivalency of the emotion labels across the corpora and allows the use of only a limited number of utterances from the corpora. Thus, a standardization of common emotion labels across emotional speech corpora is required.

This paper reports the author's attempts to confront the issues described above. First, an authentic emotional speech collection scheme was developed to confront the issue of the efficient collection of emotional speech. Then, the relationship between the two well-known types of emotion labels, i. e., emotion categories and emotion dimensions, was investigated in a cross-corpus emotion labeling experiment using two publicly available Japanese emotional speech corpora to confront the issue of standardized emotion labeling. Finally, the results of these studies are summarized in the conclusion section.

2. Collection of Authentic Emotional Speech

For the efficient collection of emotional speech, a collection scheme based on an online game task was applied, and the results were assessed in comparison with other emotional speech material. The content of this section is a rewrite of the research paper (Arimoto et al., 2012).

2.1. Recording

2.1.1. Task

To record authentic emotional expression during real-life conversations, massively multiplayer online role-playing games (MMORPGs), which are part of daily life for some Japanese university students, were adopted as tasks for our recording sessions. The effectiveness of games in evoking emotion has been proven in previous studies (Anderson and Bushman, 2001; van't Wout et al., 2006; Ravaja et al., 2008; Hazlett, 2006; Hazlett and Benedek, 2007; Tijs et al., 2008). The MMORPG used for each recording session depended on the group of players. The players in each group were allowed to select a game that more than one of them had actually played and enjoyed in their daily lives. The most popular online game was $Ragnarok\ Online$, which three groups played during recording. $Monster\ Hunter\ Frontier$ and $Red\ Stone$ were chosen by the other groups. All players were instructed to form a party and to participate together in quests (tasks in the game) while they were gaming.

To encourage the game players to talk with each other and to vocally express their emotions, an online voice chat system was adopted as a tool for communication among the players. Players of a MMORPG typically discuss their strategies for collaboratively achieving their goals in game events through a chat function provided by the MMORPG. To ensure that their emotional reactions would be reflected in their speech, the players were instructed to communicate through a voice chat system rather than the text chat function. Through the use of a voice chat system, it was expected that the players' emotional reactions to game events and expressive speech influenced by the players' internal emotional states would be observed.

2.1.2. Speakers

The speakers were 13 university students (9 males and 4 females, mean age 22 years ($SD = 1.17$)) with experience playing online games. They participated in our recording sessions as online game players. The players participated in each recording session as a group with one or two friends of the same gender. Six dialogs (five dyadic dialogs and one triad dialog) were recorded. The mean prior online gaming experience per player was 38 months ($SD = 14$), and the mean playing time per month was 33 hours ($SD = 35$).

2.1.3. Recording Environment

Figure 1 shows our recording environment. Each player in the group was located at a remote site on the campus of

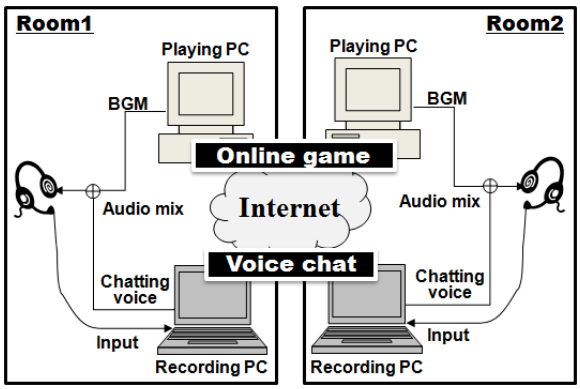

Figure 1: Recording environment.

Table 1: Number of utterances for each speaker.

Speaker	Utterances	Speaker	Utterances
01_MMK	816	04_MNN	934
01_MAD	740	04_MSY	938
02_MTN	884	05_MYH	464
02_MEM	736	05_MKK	539
02_MFM	557	06_FTY	712
03_FMA	561	06_FWA	781
03_FTY	452		
		Total	9114

Tokyo University of Technology and joined an online game together via the Internet. To make the recording environment as close as possible to the environments in which the players would usually play the game in their daily lives, a soundproof room was not used for recording. Each player sat on a chair in a classroom or on a tatami in a multipurpose space to play the game. The players put on headset microphones (Audio Technica ATH-30COM dynamic headsets) and talked with each other in a non-face-to-face environment via the Skype voice chat system. The dialogs among the players were recorded with a voice-recording system, Tapur for Skype. The speech was recorded separately at each recording site where each player was playing the game. Tapur recorded the local player's voice and a remote player's voice in different channels of a stereo sound file.

The recording time was approximately 1 hour for each group, and the total recording time was approximately 14 hours. The sound data were sampled at 48 kHz and digitized to 16 bits.

2.1.4. Segmentation and Transcription

The utterances in the recorded material were defined based on interpausal units (IPUs). Any continuous speech segment between pauses exceeding 400 ms was regarded as one utterance. The segmented utterances were orthographically transcribed into *kanji* (Chinese logograms) and *kana* (Japanese syllabograms). Jargon and special terms for online games, e. g., "bot" or "strage ("e su thi a: ru a ji" in reading)", and figures and counters were transcribed in *katakana* (angular Japanese syllabograms) as these words were heard. The following three transcription tags were prepared for laughs, coughs, and other purposes.

- {laughs},{coughs}
  Laughs, excluding utterances with laughing, and coughs.
- (?), (? (comment))
  An utterance that could not be transcribed due to noise or low sound volume.
- [comment:(comment)]
  Transcriber's comment.

Ultimately, the total number of utterances in our corpus was 9114. Table 1 shows the number of utterances for each speaker. In Table 1, the speakers are represented by speaker IDs.

2.2. Emotion Labeling

2.2.1. Speech Materials

For two speakers, 03_FMA and 02_MFM, 1009 utterances were not used in the analysis due to their low sound levels. Moreover, 1527 utterances with tags were also not used because these utterances could not be transcribed and their acoustic features could not be calculated. As a result, the total number of utterances used in the following analysis was 6578.

2.2.2. Procedure

The utterances were labeled with emotional categories in accordance with their perceived emotional information. After category labeling, the labeled utterances were rated for emotional intensity on the basis of how strongly the emotion was perceived from each utterance. Both the labelers and the raters were instructed to judge each utterance according to its acoustic characteristics, not its content.

Twenty-two labelers (14 males and 8 females) participated in the emotion labeling. Because the labeling of all 6587 utterances by each labeler would be costly and difficult, the number of utterances to be evaluated by each labeler was adjusted such that each utterance was labeled by three labelers. The labelers were instructed to choose one emotional state with which to label each utterance from ten alternatives: fear (FEA), surprise (SUR), sadness (SAD), disgust (DIS), anger (ANG), anticipation (ANT), joy (JOY), acceptance (ACC), a neutral state (NEU) with no emotion, or an utterance exhibiting an emotional state that is impossible to classify into any of the nine states above or subject to high noise or other disruption (OTH). The eight emotional states were selected with reference to the primary emotions of Plutchik's multidimensional model (Plutchik, 1980). Table 2 lists the ten emotional state classifications, their abbreviations, and their definitions. These ten definitions were presented to the labelers to give them a common understanding of each emotional state. The definitions were prepared by referring to a dictionary (Yamada et al., 2005). Each utterance was presented in a random order to each labeler to mitigate possible order effects.

Each utterance was rated for its emotional intensity by 18 raters (13 males and 5 females). Only utterances for which at least two of the three labelers agreed on one of the eight emotion labels were rated. The utterances were presented

Table 2: Abbreviations and definitions of emotional states.

State	Abbr.	Definition
Fear	FEA	Feelings of avoidance toward people or things that are harmful
Sadness	SAD	Feelings of sorrow for irrevocable consequences such as misfortune or loss
Disgust	DIS	Feelings of avoidance toward unacceptable states or acts
Anger	ANG	Feelings of irritation or annoyance with an unforgiven subject
Surprise	SUR	Feelings of being disturbed, caught off balance, or confused after experiencing unexpected events
Anticipation	ANT	Feelings of longing for a desirable eventuality or a favorable opportunity
Joy	JOY	Feelings of gladness and thankfulness indicating intense satisfaction with something
Acceptance	ACC	Feelings of active involvement in something fascinating or positive
Neutral	NEU	No feelings at all
Other	OTH	Impossible to classify into any of the nine states above, or utterances with noise, etc.

Table 3: Results of emotion labeling. The percentages were calculated by dividing the number of utterances corresponding to each emotional state by the total number of utterances. The total number of utterances was 6578.

State	Partial		Full	
	Utterances	Percent	Utterances	Percent
FEA	142	2.2	33	0.5
SAD	243	3.7	49	0.7
DIS	335	5.1	45	0.7
ANG	237	3.6	60	0.9
SUR	565	8.6	177	2.7
ANT	427	6.5	69	1.0
JOY	595	9.0	174	2.6
ACC	303	4.6	27	0.4
NEU	798	12.1	116	1.8
OTH	200	3.0	30	0.5
Total	3845	58.5	780	11.0

in a random order to each rater. The raters were instructed to rate the emotional intensity of each utterance on a five-point scale from 1 (weak) to 5 (strong).

2.3. Analysis

Two types of agreement among the three label evaluations were calculated: partial agreement (two out of three labelers agreed on one emotion) and full agreement (all three labelers agreed on one emotion). Moreover, the mean correlation coefficient among the 18 raters was calculated.

To assess the efficiency of our data collection scheme for authentic emotional speech, the number of labeled instances among our speech materials was compared with those of two other sets of speech materials. One of these consists of spontaneous pseudodialogs for angry speech classification (Ang et al., 2002), and the other is a speech database for paralinguistic information studies, the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB) (Mori, 2008). The emotion labeling rate was calculated by dividing the number of emotion labels by the total number of labels, in accordance with (Ang et al., 2002). Note that the labeling schemes for the three sets of materials are not completely the same and that the calculation was performed for the sake of comparison among them. Each utterance in the angry speech material set (Ang et al., 2002) is labeled with one of 7 emotional state labels: neutral, annoyed, frustrated, tired, amused, other, or not applicable (containing no speech data from the user). Utterances with the annoyed, frustrated, tired, amused, and other labels were regarded as emotional utterances for the comparison. The utterances in the UUDB are not labeled with single emotional states. Instead, they are rated on a seven-point scale for each of six paralinguistic information values: pleasant–unpleasant, aroused–sleepy, dominant–submissive, credible–doubtful, interested–indifferent, and positive–negative. The utterances are all associated with six paralinguistic information values; hence, a nonemotional state is never assessed. To compare the emotion labeling rates between our speech materials and the UUDB, the UUDB utterances rated with scores from 3 to 5 (weak or none) for all 6 values were regarded as nonemotional utterances, and the rest were regarded as emotional utterances. A χ^2 test was conducted to compare the emotion labeling rates among the three speech material sets.

2.4. Results

Table 3 shows the numbers of utterances exhibiting the two types of interlabeler agreement. The number of utterances with partial agreement is 3,845, and the number of utterances with full agreement is 780. The partial and full agreement rates are 58.5% (chance level: 28%) and 11.0% (chance level: 1%), respectively.

The mean correlation coefficient among the 18 raters is 0.24 (range $= -0.01$ – 0.52). The range of correlation coefficients among the 18 raters is widely spread, indicating that the criteria used to rate emotional intensity were different among the raters.

Figure 2 shows the frequency of emotion labels in each set of speech materials. The χ^2 test revealed a significant difference among the three speech material sets ($\chi^2(2) = 27659.87$, $p < 0.001$). Our speech material set has a significantly higher emotion labeling rate than the other two ($p < 0.01$, indicated by asterisks in Fig. 2).

2.5. Discussion

Quite high agreement rates were obtained for both partial and full agreement. The partial and full agreement rates are 58.5% and 11.0%, respectively, which are much higher than the chance levels for partial and full agreement (28%

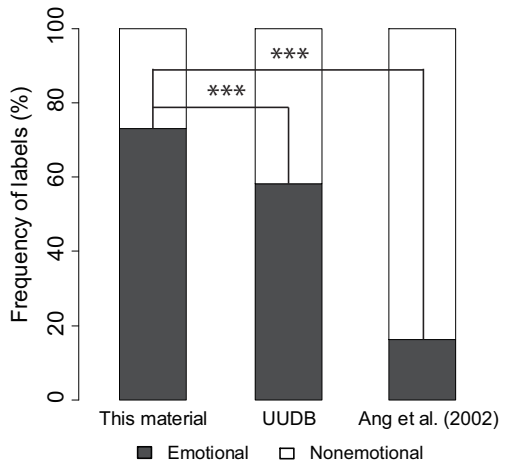

Figure 2: Frequencies of emotional labels.

and 1%, respectively). The results suggest that the labelers could perceive the same emotions from the recorded utterances. This implies that the emotional speech collected via the proposed approach is perceptually distinguishable for listeners.

The χ^2 test revealed a significant difference among the three sets of speech materials ($\chi^2(2) = 27659.87$, $p < 0.001$). Our speech material set has a significantly higher emotion labeling rate than the other two. The total number of labeling instances in our speech material set is 19,734 labels (6,578 utterances $\times$ three labelers). Among them, 14,414 labels are emotional labels corresponding to the eight types of emotional state; consequently, a very high percentage, 73.0%, of the total labeling instances have emotional labels. The total number of labeling instances in the speech material set of Ang et al. (Ang et al., 2002) is 49,553; these instances were judged by 2.62 mean labelers per utterance and include 4,904 emotional labels. The corresponding emotion labeling rate is thus quite low, 9.9%. The UUDB has 14,520 labels assigned by three labelers. Of these labels, 58.2% (8,446 labels) are emotional labels. These results imply that the proposed collection scheme can yield a relatively high percentage of emotional speech that is perceptually distinguishable by listeners.

The speech materials with emotion labels recorded via the proposed collection scheme are publicly available from the distributor, NII–SRC, as the Online Gaming Voice Chat Corpus with Emotional Label (OGVC) (Arimoto and Kawatsu, 2013).

3. Cross-corpus Emotion Labeling

To elucidate appropriate emotion labels for emotional speech and to standardize the emotion labels among several corpora, we investigated the relationship between two well-known types of emotion labels, i. e., emotion categories and emotion dimensions. Using two publicly available Japanese dialog speech corpora with emotion labels, we conducted cross-corpus emotion labeling to label the utterances in the two corpora with both emotion category labels and emotion dimension labels. The content of this section is a rewrite of the conference paper (Arimoto and Mori, 2017).

3.1. Speech Materials

Two publicly available Japanese dialog speech corpora were used for this research: the OGVC (Arimoto and Kawatsu, 2013) and the UUDB (Mori, 2008).

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

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

The emotional state corresponding to each utterance is evaluated on a seven-point scale for each dimension. On 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 in this experiment.

3.2. Procedure

The two corpora used in this study have different types of emotion labels; consequently, they cannot be used together for any research in their original forms. Therefore, in this experiment, the emotion labels included in the original corpora were discarded, and all utterances in both corpora were newly labeled with emotion categories and emotion dimensions to obtain common emotion labels across the two corpora.

Three qualified labelers, selected via a previously performed labeler screening process, performed the cross-corpus emotion labeling. The mean age of the three labelers was 22 years ($SD = 0.82$).

The emotion labeling frameworks for both emotion category labeling and emotion dimension labeling were the same as those used in the construction of the two original corpora. For emotion category labeling, the labelers were instructed to choose one of 10 categories (JOY, ACC, FEA, SUR, SAD, DIS, ANG, ANT, NEU, and OTH) for each utterance. The ground-truth label for each utterance was determined by majority vote among the labelers. For emotion dimension labeling, the labelers were instructed to rate each of the six emotion dimensions on a seven-point scale for each utterance. The ground-truth label for each emotion dimension for each utterance was defined as the mean score among the labelers. Each labeler performed both the emotion category and emotion dimension labeling tasks. The emotion dimension labeling task preceded the emotion category labeling task.

Table 4: The number of utterances in 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

Each labeler evaluated a total of 11,418 utterances from the OGVC and the UUDB (6,578 from the OGVC and 4,840 from the UUDB). The 11,418 utterances were randomly separated into blocks. The cross-corpus emotion labeling was performed in 104 blocks for 11,418 utterances $\times$ 2 types of labeling (category and dimension).

3.3. Analysis

To assess the independence of each emotion category from the others in an n-dimensional emotional space, equivalence tests between two n-dimensional Gaussian mixture models (GMMs) were conducted. For each pair of 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 GMMs. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ denote the subdatasets belonging to E_1 and E_2, respectively, and N_1 and N_2 denote the respective data sizes. 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 that 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 that differs from M_1.

The null hypothesis can be tested using a parametric bootstrap likelihood ratio test, in which 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 identical models on random samples with a data size of N_1+N_2) and the alternative model (M_1 and M_2 are trained separately on random samples with a data size of N_1 and random samples with a data size of N_2, respectively) is estimated via random sampling under H_0. If the difference of the deviances between the null model (identical GMMs trained on $\mathbf{x}_1 + \mathbf{x}_2$) and the alternative model (GMMs trained separately on $\mathbf{x}_1$ and $\mathbf{x}_2$) falls into the critical region ($\alpha = 5\%$), then the null hypothesis is rejected, and the two emotion categories are considered to be independently distributed in the n-dimensional emotional space. Such likelihood ratio tests were conducted for all combinations of the nine emotion categories.

3.4. Results

Table 4 shows the number of utterances in each emotional category identified as a result of the emotion category labeling process. The total number of utterances for which two out of the three labelers agreed on one emotion label is 5,797 (3,446 for the OGVC and 2,351 for the UUDB), corresponding to 51% of the total utterances subjected to cross-corpus labeling (52% of the OGVC utterances and 49% of the UUDB utterances). The emotions assigned to the highest numbers of utterances, in descending order, are ACC, DIS, JOY and SAD. Following emotion category labeling, these 5,797 utterances were used in the analysis of the mapping of the emotion categories to n-dimensional emotional spaces.

Table 5: Differences in deviances between emotion categories mapped to a 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$*$

Figure 3 shows the distributions of the emotion categories in the two-dimensional emotional spaces of arousal vs. pleasantness, dominance vs. pleasantness, and dominance vs. arousal. Table 5 shows the differences in the deviances between the emotion categories when mapped to the corresponding three-dimensional emotional space. The asterisks in Table 5 indicate the combinations of emotion categories for which the hypothesis H_0 is rejected and the hypothesis H_1 is accepted ($p < 0.05$). For many combinations of emotion categories, H_0 is rejected; H_0 was not rejected in only three tests, namely, for NEU when testing with SUR, SAD, and DIS.

3.5. Discussion

In the pleasantness vs. arousal space shown in the left panel of Fig. 3, JOY (the solid red line in Fig. 3) is placed in the upper right quadrant, corresponding to high arousal and high pleasantness; SUR (dashed green line) corresponds to high arousal; SAD (solid green line) corresponds to low arousal; and ANG (solid blue line) lies in the upper left, corresponding to high arousal and low pleasantness. These distributions are similar to Russell's circumplex model (Russell, 1980). The results also show that NEU (solid purple line) lies near 4 on the pleasantness axis but between 2 and 4 on both the arousal and dominance axes. NEU is generally considered to be an emotionally neutral state, which should correspond to a score of 4 in any emotion dimension. However, our results imply that neutral utterances are neutral in the pleasantness dimension but are not necessarily neutral in the other dimensions.

The results of the likelihood ratio tests on the distributions of the emotion categories in the three-dimensional emotional space suggest that all pairs of emotion categories except NEU–SUR, NEU–SAD, and NEU–DIS exhibit significant differences between each other ($p < 0.05$). In other words, all emotion categories except NEU are independent of each other. This finding suggests that the information of the eight emotion categories (JOY, ACC, FEA, SUR, SAD, DIS, ANG, and ANT) is not lost even in the emotion di-

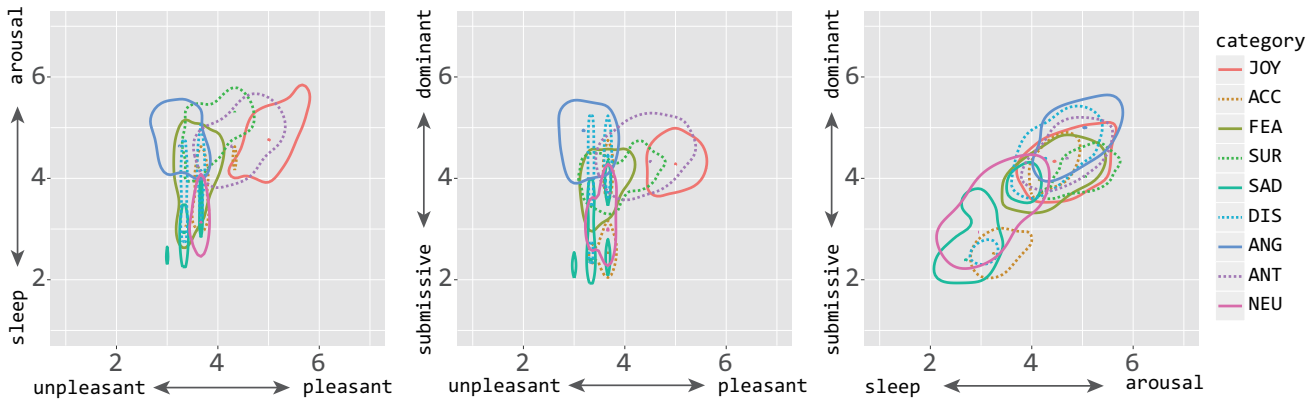

Figure 3: Distributions of emotion categories in two-dimensional emotional spaces.

mension representation.

4. Conclusions

For the efficient collection of emotional speech, a collection scheme based on an online game task and a voice chat system was developed, and its results were assessed by comparison with other emotional speech materials. A χ^2 test revealed that by using the proposed collection scheme, emotionally expressive speech can be efficiently collected.

To elucidate appropriate emotion labels for emotional speech and to commonize emotion labels among several corpora, we first studied the relationship between emotion categories and emotion dimensions. Using two Japanese dialog speech corpora with emotion labels, cross-corpus emotion labeling was conducted to label the utterances in the two corpora with both emotion category labels and emotion dimension labels. Then, likelihood ratio tests were conducted to assess the independence of each emotion category from the others in a three-dimensional emotional space.

The tests revealed that all pairs of emotion categories except neutral–surprise, neutral–sadness, and neutral–disgust exhibit significant differences between each other. Thus, all emotion categories except neutral are independent of each other in the dimensional emotional space.

These results suggest the surprising conclusion that the information of the eight emotion categories, including joy, acceptance, fear, surprise, sadness, disgust, anger, and anticipation, is not lost even in the emotion dimension representation. However, future research with other speech corpora in different languages may yield different results, because emotion perception heavily depends on language, culture and social norms. The universal standardization of emotion labeling can be accomplished only after examining the linguistic differences, cultural differences, and social differences that must be encompassed by standardized emotion labels.

5. Acknowledgments

This work was supported by a TATEISHI Science and Technology Foundation Grant for Research (A) and by JSPS KAKENHI Grant Number 17K00160. I would like to express my sincere appreciation to my collaborators, Dr. Hiromi Kawatsu from IBM and Prof. Hiroki Mori from Utsunomiya University.

6. Bibliographical References

Anderson, C. a. and Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: a meta-analytic review of the scientific literature. *Psychological science*, 12(5):353–9, sep.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A. (2002). Prosody-Based Automatic Detection Of Annoyance And Frustration In Human-Computer Dialog. In *Proceedings of ICSLP 2002*, pages 2037–2040. in Proc. ICSLP 2002.

Arimoto, Y. and Mori, H. (2017). Emotion category mapping to emotional space by cross-corpus emotion labeling. In *Proceedings of Interspeech 2017*, pages 3276–3280.

Arimoto, Y., Ohno, S., and Iida, H. (2007). An Estimation Method of Degree of Speaker's Anger Emotion with Acoustic and Linguistic Features. *Journal of natural language processing*, 14(3):147–163. (in Japanese).

Arimoto, Y., Kawatsu, H., Ohno, S., and Iida, H. (2008). Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems. In *Proceedings of Interspeech 2008*, pages 322–325.

Arimoto, Y., Kawatsu, H., Ohno, S., and Iida, H. (2012). Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment. *Acoustical Science and Technology*, 33(6):359–369.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636.

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2003). How to find trouble in communication. *Speech Communication*, 40(1-2):117–143, apr.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., and Kessous, L. (2011). Whodunnit - Searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language*, 25(1):4–28, jan.

Campbell, N. (2004). Speech & Expression ; the Value of a Longitudinal Corpus The JST ESP corpus. In *LREC 2004*.

Cowie, R. (2009). Perceiving emotion: towards a realis-

tic understanding of the task. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1535):3515–25, dec.

Devillers, L., Vidrascu, L., and Bp, L.-c. (2006). Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs. In *Interspeech 2006.*, pages 801–804.

Douglas-Cowie, E. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, apr.

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

Engberg, I. S., Hansen, A. V., Andersen, O., and Dalsgaard, P. (1997). Design, Recording and Verification of a Danish Emotional Speech Database. In *Proceedings of Eurospeech 1997*, volume 4, pages 1695 – 1698.

Hazlett, R. and Benedek, J. (2007). Measuring emotional valence to understand the user's experience of software. *International Journal of Human-Computer Studies*, 65(4):306–314, apr.

Hazlett, R. L. (2006). Measuring emotional valence during interactive experiences. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, pages 1023–1026, New York, New York, USA, apr. ACM Press.

Itoh, K. (1986). A basic study on voice sound involving emotion. III. Non-stationary analysis of single vowel [e]. *The Japanese journal of ergonomics*, 22(4):211–217.

Jürgens, R., Hammerschmidt, K., and Fischer, J. (2011). Authentic and play-acted vocal emotion expressions reveal acoustic differences. *Frontiers in psychology*, 2(July):180, jan.

Kitahara, Y. (1988). Prosodic components of speech in the expression of emotions. *The Journal of the Acoustical Society of America*, 84(S1):S98–S99, nov.

Litman, D. and Forbesriley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590, may.

Mori, H., Satake, T., Nakamura, M., and Kasuya, H. (2011). Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, 53(1):36–50, aug.

Plutchik, R. (1980). *Emotions: A psychoevolutionary synthesis*. Harper & Row, New York.

Ravaja, N., Turpeinen, M., Saari, T., Puttonen, S., and Keltikangas-Järvinen, L. (2008). The psychophysiology of James Bond: phasic emotional responses to violent video game events. *Emotion (Washington, D.C.)*, 8(1):114–20, feb.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

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

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-Corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Schuller, B., Zhang, Z., Weninger, F., and Burkhardt, F. (2012). Synthesized speech for model training in cross-corpus recognition of human emotion. *International Journal of Speech Technology*, 15(3):313–323.

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

Tijs, T., Brokken, D., and Ijsselsteijn, W. (2008). Creating an Emotionally Adaptive Game. In S M Stevens et al., editors, *Proceedings of the 7th International Conference on Entertainment Computing*, volume 5309 of *LNCS 5309*, pages 122–133. Springer-Verlag.

van't Wout, M., Kahn, R. S., Sanfey, A. G., and Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 169(4):564–8, mar.

Williams, C. E. (1972). Emotions and Speech: Some Acoustical Correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250, oct.

Yamada, T., Shibata, T., Kuramochi, Y., and Yamada, A. (2005). *Shin meikai kokugo jiten*. Sanseido, Tokyo, 6 edition. (in Japanese).

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58.

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

Zong, Y., Zheng, W., Zhang, T., and Huang, X. (2016). Cross-Corpus Speech Emotion Recognition Based on Domain-Adaptive Least-Squares Regression. *IEEE Signal Processing Letters*, 23(5):585–589, may.

7. Language Resource References

Arimoto, Yoshiko and Kawatsu, Hiromi. (2013). *Online gaming voice chat corpus with emotional label (OGVC)*. Speech Resource Consortium, National Institute of Informatics, ISLRN 648-310-192-037-7.

Mori, Hiroki. (2008). *Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB)*. Speech Resource Consortium, National Institute of Informatics.