Correlation Analysis between Subjectively Annotated Emotions and Objectively Annotated Emotions

Shota Saitou, Kazuya Mera, Yoshiaki Kurosawa, and Toshiyuki Takezawa

Abstract— **There is no research for analyzing the relationship between subjectively annotated emotions and objectively annotated emotions despite the fact that a lot of research uses objective emotion labels as subjective emotions. In this study, we collect natural emotional voices, and the speaker and others annotate the intensity of six emotions {anger, dislike, fear, happiness, sadness, and surprise} to the voices. The correlation diagrams between subjectively and objectively annotated emotions indicate that there is a small relationship between subjective and objective emotions. The experimental results using support vector regression reveal that learning subjective emotions is much more difficult than learning objective emotions. Furthermore, happiness, sadness, and surprise are comparatively learned better than dislike, anger, and fear.**

Index Terms—**self-reported emotion, objectively perceived emotion, emotional voice, machine learning**

I. Introduction

VARIOUS machine learning models such as support vector machines and deep neural networks are currently used to estimate human emotions. Such supervised learning models require training data, which should be labeled with an emotion class (for classification) or by the intensity of the emotion (for regression). One might ask the question, "what is annotated as the emotion label?"

Much research is concerned with the accurate detection of human emotions aroused at any given time; however, it is difficult to annotate the emotions. Therefore, various emotion labels are substituted for the aroused emotions. Fig. 1 shows various emotion labels substituted for an emotional reaction.

self-reported emotion: The speaker annotates his/her own aroused emotion at its utterance. Although the annotation seems to be the same as the aroused emotion because the annotator is the speaker him- or herself, it is difficult to correctly recall the aroused emotion later. On the other hand, thinking about how to annotate an emotion may interfere with its natural arousal if the speaker utters and annotates the emotion at the same time.

Manuscript received Dec 19, 2018; revised Jan 4, 2019. This work was supported by the Center of Innovation Program from Japan Science and Technology Agency, JST.

Kazuya Mera, Shota Saitou, Yoshiaki Kurosawa, and Toshiyuki Takezawa are with the Graduate School of Information Sciences, Hiroshima City University, Hiroshima, Japan
(e-mail: {mera, kurosawa, takezawa}@hiroshima-cu.ac.jp).

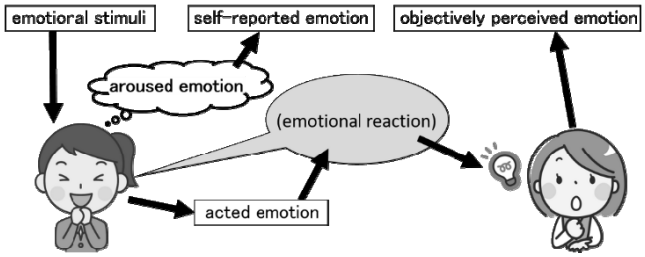

Fig. 1. Various emotion labels

emotional stimuli: Some stimuli can induce emotional reactions. An emotional reaction to a stimulus is recognized as a reaction to the emotion rated to the stimulus. This method is commonly used in psychological fields. However, not all people react to the stimulus with the annotated emotion.

acted emotion: In order to collect many emotional reactions of the required emotion, it is comparatively easy to make participants act the emotional reactions. However, such reactions often have different features from natural emotional reactions because some features may be exaggerated or suppressed unnaturally.

objectively perceived emotion: Certain communication, unlike that between a counselor and a client, does not require accurate estimation of emotions. The ability to estimate emotion on the same level as ordinary people is often enough for daily use. Furthermore, collecting a large amount of annotated data for the required emotional reactions is easy, and the annotation results are relatively reliable because the emotion labels are statistically defined.

Despite discrepancies between emotions estimated by another person and the actual emotions aroused in a speaker, objectively perceived emotions are often dealt with as aroused emotions.

In this study, we construct an emotional voice database labeled with both self-reported and objectively perceived emotions. Self-reported emotions are rated for the intensity of the aroused emotion by the speaker him- or herself. Objectively perceived emotions are rated for the likelihood and intensity calculated from the annotated results by 47 annotators. Then, the correlation between the self-reported emotions and the objectively perceived emotions is analyzed. Furthermore, the aptitudes of these emotion labels for machine learning are examined.

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. Previous Studies of Emotional Voice Databases

In order to investigate the differences of acoustic features among emotions or to use them for machine learning training data, various emotional voice databases have been developed.

FAU Aibo Emotion Corpus [1] consists of nine hours of German speech of 51 children at 10-13 years of age interacting with Sony's pet robot Aibo. The audio recordings of the children have been segmented manually into small, syntactically meaningful 'chunks' using syntactic-prosodic criteria. The data are annotated with 11 emotion categories by five human labelers on the word level. Online Game Voice Chat Corpus with Emotional Label (OGVC) [2] collects naturalistic emotional speech during a spontaneous Japanese dialog. To encourage speakers to experience and express their natural and vivid emotions, a massively multiplayer online role-playing game (MMORPG) was used. Each utterance was labeled by three labelers, and the labelers had to choose one emotional state from fear, surprise, sadness, disgust, anger, anticipation, joy, acceptance, neutral, and other. Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB) [3] collects conversations while participants perform the "4-frame cartoon sorting task." The perceived emotional states were annotated with six abstract dimensions: pleasant-unpleasant, aroused-sleepy, dominant- submissive, credible-doubtful, interested-indifferent, and positive-negative. Although these databases collect natural emotional voices, the voices are annotated by annotators who just listen to the voices. As described in Chapter 1, ordinary people cannot always accurately estimate the emotion aroused in the speaker.

On the other hand, OGVC [2], Surrey Audio-Visual Expressed Emotion (SAVEE) Database [4], and Berlin Database of Emotional Speech (Emo-DB) [5] collect acted emotional voices and the acted emotions are labeled to the voices. Acted emotional reactions may also contain an actor's actual emotion because some actors can arouse pseudo-emotions when they act emotional reactions. For example, Stanislavski's system [6] proposes a method for arousing emotion as a character by imagining the experience, background, and thought of the character in detail.

Furthermore, databases of emotional stimuli have also been developed. For example, International Affective Digital Sounds (IADS) [7] and International Affective Picture System (IAPS) [8] provide a set of emotional stimuli, and the stimuli are rated on the dimensions of pleasure, arousal, and dominance because an emotional reaction by a stimulus is recognized as that of the emotion rated to the stimulus in psychological fields.

However, we could not find databases of emotional reactions that labeled genuine "aroused emotions." One of the reasons why it is difficult for others to accurately estimate a speaker's emotion and to think of an annotation may be that uttering and annotating an emotion at the same time interferes with the naturally aroused emotion.

Therefore, we collect emotional voices labeled as "self-reported emotions" instead of "aroused emotions" and as "objectively perceived emotions" by others.

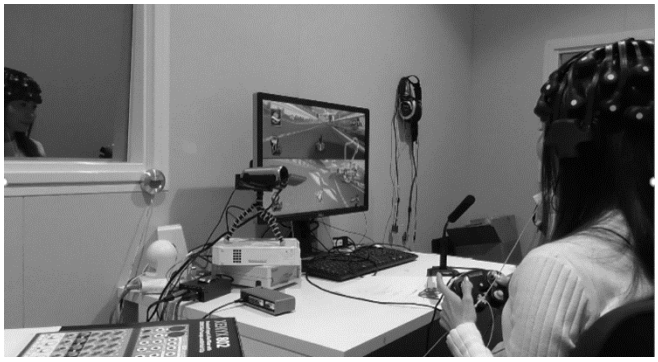

Fig. 2. Emotional voice recording environment

TABLE I
Personality Type Diagnosis Results of Participants

Personality Type	Number of Participants
Balanced	0
Introvert	6
Matured	4

III. Emotional Voice Database with Subjectively and Objectively Annotated Emotion Labels

A. Collecting Emotional Voices

To ask the speakers themselves to annotate their emotions using a "self-reported emotion" label, we collected emotional voices and made the speakers annotate their emotions to their own utterances right after recording their voices. To encourage speakers to experience and express their natural and vivid emotions, a go-kart-style racing video game was used just like with OGVC [2]. Ten participants (21-24 year-old university students) joined the experiment. Although the experiment collects not only voice but also facial expressions and brain waves at the same time, this research uses the emotional voices only. The experimental procedure is as follows:

(1) Staff pairs participants.
(2) Staff puts an earphone and an electroencephalogram (EEG) on each participant.
(3) Staff sets a video camera and a high directional microphone for each participant.
(4) Staff sets another video camera to record the game display and participants' voices simultaneously.
(5) Participants play the video game for ten minutes, and conversation voices and other data are recorded by using (2), (3), and (4) while the game is played.
(6) Participants annotate all of their voices with the emotional labels by watching (4).

In addition, the participants answered a "Big-Five" personality questionnaire (NEO-FFI), and they were classified into the three personality types (Balanced, Introvert, and Matured) proposed by Machizawa [9]. Table I shows the number of participants classified into the three personality types.

1,367 emotional voices were collected by the recording process. Section III-B explains the annotation process (6) in detail.

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

B. Annotation of Self-Reported Emotion Labels

It seems better to collect emotional reactions and have participants annotate their emotions at the same time to know exactly what emotion is aroused at that time. However, the annotation process may interfere with arousing naturalistic emotions. On the other hand, participants often forget their own emotions if they do not annotate their voices soon after utterance.

Therefore, the participants annotated the emotion labels to their own voices just after the recording process. When they annotate their voices, they listen to their voices and watch the game display at that time to enable recollection of aroused emotions.

The participants annotated all of their own voices with the intensities of Ekman's six basic emotions {anger, dislike, fear, happiness, sadness, and surprise} [10]. The intensities of emotions were annotated using the visual analogue scale (VAS) [11], and the intensities were translated into the range of [0, 100]. Table II shows the number of annotated intensities for each emotion.

C. Annotation of Perceived Emotion Labels

To calculate objectively perceived emotion, 47 annotators (18-24 year-old university students, 36 males and 11 females, except for the speaker of the annotated voice) estimated the speaker's emotion and annotated the emotion labels to 1,367 collected voices. The annotators judged the intensities of six emotions {anger, dislike, fear, happiness, sadness, and surprise} for each voice by the following four grades:

0: Speaker does not arouse *X*.
1: Speaker arouses *X* a little.
2: Speaker arouses *X*.
3: Speaker strongly arouses *X*.

where *X* is the corresponding emotion.

This database annotates two kinds of perceived emotion labels: "objective likelihood (Obj_Like)" and "objective intensity (Obj_Int)." The likelihood and intensity of emotion are calculated from annotated intensities by the annotators. Obj_Like is defined as the rate of annotators who annotated that the speaker more or less arouses the emotion, i.e., the intensity of the emotion is 1, 2, or 3. Obj_Int is defined as an average of annotated intensities of the emotion. Obj_Like and Obj_Int are calculated by (1) and (2), respectively.

$$Obj\_Like = \frac{\sum_{i=1}^{3} Emo_i}{\sum_{i=0}^{3} Emo_i} \quad (1)$$

$$Obj\_Int = \frac{Emo_0 \times 0 + Emo_1 \times 1 + Emo_2 \times 2 + Emo_3 \times 3}{\sum_{i=0}^{3} Emo_i} \quad (2)$$

where Emo_0, Emo_1, Emo_2, and Emo_3 are the number of annotators who annotated the intensity of the emotion as 0, 1, 2, and 3, respectively.

TABLE II
AMOUNT OF DATA FOR EACH INTENSITY OF SELF-REPORTED EMOTION

	Intensity of Self-Reported Emotion						sum
	0	[1,20]	[21,40]	[41,60]	[61,80]	[81,100]	
anger	1,029	93	65	39	75	66	1,367
dislike	968	94	111	102	63	29	1,367
fear	1,107	66	50	54	48	42	1,367
happiness	682	79	117	156	153	180	1,367
sadness	944	71	73	81	80	118	1,367
surprise	888	107	84	79	122	87	1,367

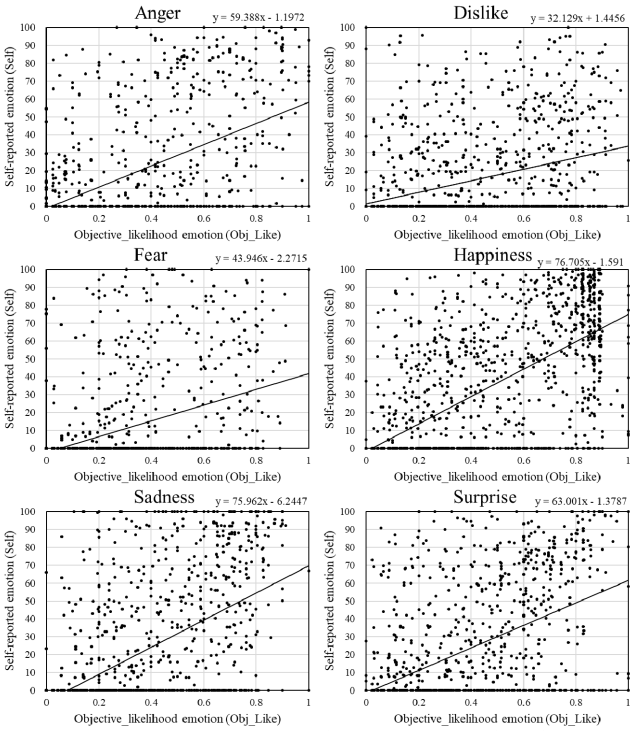

Fig. 3. Correlation diagrams between Self and Obj_Like

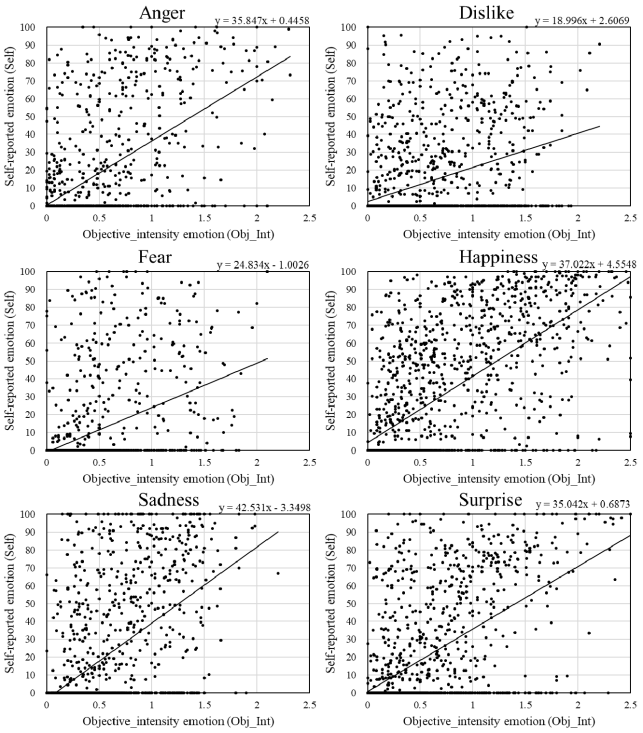

Fig. 4. Correlation diagrams between Self and Obj_Int

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

D. Correlation between Self-Reported Emotions and Objectively Annotated Emotions

This section analyzes the correlations between subjectively and objectively annotated emotions by using the emotional voice database constructed in Chapter III. In this paper, a self-reported emotion label is regarded as a subjective emotion label.

The correlation diagrams between self-reported emotions (Self) and objectively annotated emotions (likelihood (Obj_like) and intensity (Obj_Int)) are shown in Fig. 3 and 4. The diagrams indicate that there is a small relationship between subjective and objective emotions. Even out of those, happiness, sadness, and surprise are relatively well correlated between self-reported emotions and objective emotions. Then, the emotional voices for which the speaker felt low intensity but the others annotated as high intensity are not much (lower-right area of the graphs) as shown in Fig 4. However, the voices for which the speaker felt high intensity but the others annotated as low intensity often appeared (higher-left area of the graphs). These results indicate that subjective feeling and objective estimation do not correlate well against our expectations.

IV. EXPERIMENTATION

In previous research for estimating human emotions from voices, objectively annotated emotion labels are used as supervisory signals for supervised machine learning. However, Chapter III indicates that objectively annotated emotion labels are less correlated with subjectively annotated emotion labels.

In this chapter, we construct three types of learning machines for regression that learn self-reported emotion labels, objective-likelihood emotion labels, and objective-intensity emotion labels, respectively.

A. Experimentation Method

In this experiment, support vector regression (SVR) [12] is utilized as a regression machine. The input signals are the acoustic features of the voice, and the supervisory signal is the annotated emotion label. OpenSMILE [13] calculates the acoustic features of the voice. In this experiment, "the openSMILE 'emobase2010' reference set" is used for input signals. The set contains 1,582 features relative to loudness, mel-frequency cepstral coefficients, fundamental frequency, voicing probability, jitter, shimmer, number of pitch onsets, and length of the input. The input signals are normalized before machine learning. The SVR machines use the radial basis function (RBF) kernel.

Three experimentations were carried out in which the supervisory signal is "self-reported emotion," "objective-likelihood emotion," and "objective-intensity emotion." The amount of input data is 1,367, and the regression machines are evaluated by leave-one-out cross validation.

B. Experimental Results

Fig. 5, 6, and 7 shows the correlation diagrams between the actual and predicted value of each emotion label. The number of the plots in Fig 5 seems less than that in Fig. 6 and 7 because more data are plotted on the y-axis; in other words, many self-reported emotion values of the training data are 0 as shown in Table II.

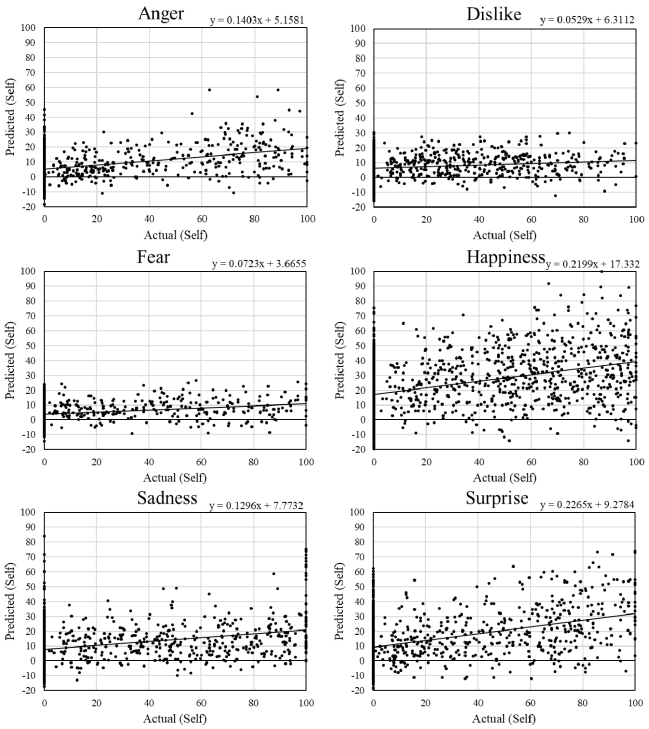

Fig. 5. Correlation diagrams of SVR results for self-reported emotions

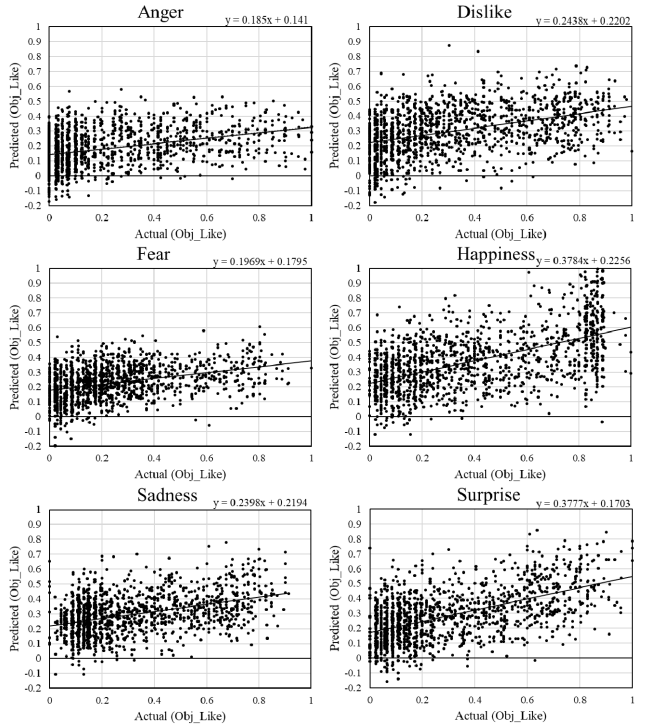

Fig. 6. Correlation diagrams of SVR results for objective-likelihood

By comparing Fig. 5 and Fig. 6 and 7, learning subjective emotion labels is much more difficult than learning both of the objective emotion labels despite using the same training data and algorithm.

Next, which emotion class was more suitable for machine learning was analyzed. Furthermore, we applied SVR to the data that were classified into each personality group and for individuals, too. Table III is the results of the Pearson Correlation Coefficient (PCC) between the predicted and actual values for each data group. Table IV is the results of Spearman's rank correlation coefficient (Spearman's rho) between the predicted and actual value for each data group.

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

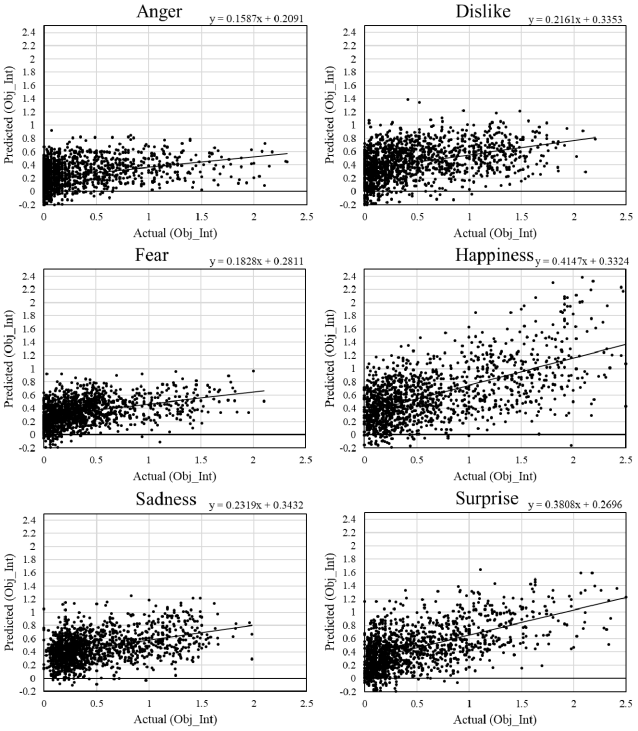

Fig. 7. Correlation diagrams of SVR result for objective-intensity

The name of data group "P$n(t)$" indicates that the data group consists of the emotional voices of the nth participant and the participant is classified into personality type t in Table III and IV. The PCC and Spearman's rho test the null hypothesis that $\rho = 0$.

Table III and IV indicate that happiness, sadness, and surprise can be learned well by SVR. On the contrary, it was difficult to predict the intensities of anger, dislike, and fear. However, no significant differences among personality types (Introvert and Matured) were found. On the contrary, the performance among speakers was quite different. The voices of P3, P4, P9, and P10 were learned well, but those of P1, P5, and P8 were not learned well.

C. Considerations

The learning machine for subjectively annotated emotions could not learn as well as that for objectively annotated emotions. One of the reasons is that the learning machine considers only acoustic features just like objective annotation, despite the speaker annotating his or her own emotions using multimodal information. The performance of the learning machine will be improved by using further modalities such as facial expressions and physiological signals.

Nevertheless, happiness, sadness, and surprise were comparatively well learned because the voices with strong happiness, sadness, and surprise (i.e., the intensities of self-reported emotions were more than 60) could be collected more than the other emotions as shown in Table II. If enough emotional voices with strong anger, dislike, and fear are prepared, the performance of the learning machine will be improved. Such emotional voices can be collected if emotional stimuli to arouse a lot of negative emotions are used on participants, provided that the mental conditions of the participants are taken into account.

TABLE III
PCC RESULTS OF SVR OUTPUT (SELF-REPORTED EMOTIONS)

Data Group	Emotion Type						Amount of Data
	Anger	Dislike	Fear	Happiness	Sadness	Surprise	
All	**0.38****	**0.16****	**0.25****	**0.40****	**0.32****	**0.42****	**1,367**
Introvert	**0.38****	**0.13****	**0.28****	**0.43****	**0.32****	**0.41****	**909**
Matured	**0.31****	**0.06****	**0.31****	**0.29****	**0.34****	**0.40****	**458**
P1 (Intro)	0.14	0.12	0.11	**0.36****	**0.20****	**0.20****	**233**
P2 (Intro)	-0.05	0.04	0.17	**0.53****	**0.50****	**0.48****	**131**
P3 (Intro)	**0.62****	**0.34****	-0.08	**0.62****	0.09	**0.44****	**184**
P4 (Intro)	0.10	**0.28****	**0.48****	**0.63****	**0.32****	0.02	**133**
P5 (Intro)	0.07	0.01	**0.28****	0.12	-0.02	**0.25****	**110**
P6 (Intro)	0.15	**0.22***	0.04	**0.34****	**0.36****	**0.44****	**97**
P7 (Matur)	**0.22****	-0.01	**0.38****	**0.26****	0.06	**0.42****	**160**
P8 (Matur)	0.03	0.09	0.03	-0.03	0.07	**0.32****	**73**
P9 (Matur)	0.21	**0.28****	**0.55****	**0.33****	**0.35****	**0.39****	**105**
P10 (Matur)	**0.36****	**0.27****	**0.28****	**0.51****	**0.42****	**0.25****	**141**

*: p<0.05, **: p<0.01

TABLE IV
SPEARMAN'S RHO OF SVR OUTPUT (SELF-REPORTED EMOTIONS)

Data Group	Emotion Type						Amount of Data
	Anger	Dislike	Fear	Happiness	Sadness	Surprise	
All	**0.29****	**0.17****	**0.26****	**0.39****	**0.28****	**0.36****	**1,367**
Introvert	**0.22****	**0.13****	**0.15****	**0.40****	**0.30****	**0.38****	**909**
Matured	**0.27****	**0.09***	**0.20****	**0.34****	**0.26****	**0.31****	**458**
P1 (Intro)	**0.14***	0.10	0.12	**0.34****	**0.18****	**0.20****	**233**
P2 (Intro)	0.03	0.09	**0.20***	**0.56****	**0.48****	**0.51****	**131**
P3 (Intro)	**0.45****	**0.38****	0.04	**0.55****	**0.16***	**0.44****	**184**
P4 (Intro)	0.14	**0.34****	**0.50****	**0.67****	**0.40****	0.01	**133**
P5 (Intro)	0.07	0.05	**0.24***	0.14	-0.01	**0.27****	**110**
P6 (Intro)	**0.25***	**0.25***	0.10	**0.34****	**0.35****	**0.43****	**97**
P7 (Matur)	**0.19***	0.03	**0.33****	**0.30****	0.10	**0.37****	**160**
P8 (Matur)	0.12	0.10	0.01	-0.01	-0.03	0.22	**73**
P9 (Matur)	**0.26***	**0.27***	**0.49****	**0.37****	**0.32****	**0.41****	**105**
P10 (Matur)	**0.34***	**0.34****	**0.22****	**0.49****	**0.46****	**0.28****	**141**

*: p<0.05, **: p<0.01

V. CONCLUSION

In this paper, we constructed an emotional voice database that labels both self-reported and objectively perceived emotions. The collected 1,367 emotional voices were annotated by the speakers themselves and the other 47 annotators. The speakers annotated the subjective emotion labels just after the recording process by listening to their voices and watching the recording scene. The annotators judged the intensities of six emotions {anger, dislike, fear, happiness, sadness, and surprise} for each voice by four grades of intensities. Objective-likelihood and objective-intensity labels were calculated from the annotation results.

The correlation diagrams indicated that there is a small relationship between subjective and objective emotions. Then, three types of machine learning tasks for regression that use self-reported, objective-likelihood, and objective-intensity emotion labels, respectively, are conducted. The experimental results revealed that learning subjective emotions was much more difficult than learning objective emotions. Furthermore, happiness, sadness, and surprise were comparatively learned better than the other emotions.

In future work, further modalities such as facial expressions and physiological signals will be used for input features. Furthermore, we have to collect emotional voices with strong anger, dislike, and fear while taking into account the mental conditions of the participants.

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

REFERENCES

[1] A. Batliner, C. Hacker, S. Steidl, E. Noth, S. D'Arcy, M. Russell, and M. Wong, "You stupid tin box' - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus.," in *Proc. of LREC 2004*, pp. 171-174.

[2] 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.

[3] 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.

[4] S. Haq and P. J. B. Jackson, "Multimodal emotion recognition," in, *Machine Audition: Principles, Algorithms and Systems*, W. Wang (Ed.), IGI Global Press, 2010, chapter 17, pp. 398-423.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Procs. of Interspeech 2005*, pp. 3-6.

[6] K. Stanislavski, An actor's work: A student's diary, Trans. and Ed. J. Benedetti, London and New York: Routledge, 2008.

[7] M. M. Bradley and P. J. Lang, "The international affective digitized sounds (2nd Edition; IADS-2): Affective ratings of sounds and instruction manual," *Technical report B-3*, University of Florida, Gainesville, Fl. 2007.

[8] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," *Technical report B-3*, University of Florida, Gainesville, Fl. 2008.

[9] M. G. Machizawa, N. Kanayama, K. Makita, T. Sasaoka, G. Lisi, and S. Yamawaki, "A simplified multi-axis affective and cognitive decoded neurofeedback system for anticipation of excitement," in *Proc. of Real-Time Functional Imaging and Neurofeedback conference (RTFIN)*, Nara, Japan, Nov. 2017.

[10] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial expressions*, Malor Books, 2003.

[11] M. H. S. Hayes and D. G. Patterson, "Experimental development of the graphic rating method," *Psychological Bulletin*, vol. 18, pp. 98-99, 1921.

[12] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems 9*, MOT Press, 1996, pp. 155-161.

[13] F. Eyben, F. Weninger, F. Gross, B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. of ACM Multimedia (MM)*, Barcelona, Spain, ACM, pp. 835-838, October, 2013.

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)