

ACOUSTICAL LETTER

Assessment of spontaneous emotional speech database toward emotion recognition: Intensity and similarity of perceived emotion from spontaneously expressed emotional speech

Yoshiko Arimoto*,†, Sumio Ohno and Hitoshi Iida

Tokyo University of Technology,
1404-1 Katakura, Hachioji, 192–0982 Japan

(*Received 14 April 2010, Accepted for publication 8 July 2010*)

Keywords: Emotional speech database, Spoken dialog, Perceived emotion, MMORPG, Voice chat
PACS number: 43.66.Yw [doi:10.1250/ast.32.26]

1. Introduction

With the great advance in automatic speech recognition, demands for automatic emotion recognition are expanding in order to facilitate smoother human-computer or computer-aided communication. To realize emotion recognition from real-life conversation, a spontaneous emotional speech database is indispensable. There are several spontaneous speech corpora for speech recognition or even paralinguistic information recognition [1,2]. However, there are quite a few spontaneous Japanese emotional speech corpora focused on emotion recognition.

In this letter, we assess our spontaneous emotional speech database comprising utterances from dialogs among online game players chatting via a voice chat system, and discuss the component of space of perceived emotion from the viewpoint of intensity and similarity.

2. Spontaneous emotional speech database

2.1. Aim and nature of our database

Many studies of emotional speech were conducted with a database comprising utterances that a professional or amateur performer spoke in a typical expression of a specific emotion [3]. These acted utterances were suitable for emotional speech synthesis because they are expressive and easy to convey the speaker's emotion to listeners. However, acted utterances were inadequate for emotional speech recognition because these utterances were more unnatural and more exaggerated than utterances in our daily life. Utterances of spontaneous emotion should be the target for emotion recognition.

To collect a wide variety of spontaneously expressed emotional speech, dialogs among online game players were recorded. Only a massively multiplayer online role-playing game (MMORPG) was adopted for our recording to force players to talk a lot to discuss plans and strategies to advance the game.

2.2. Game players

The players were 13 university students (9 males and 4 females) with experience in the online game. Their average playing time per month was 33 hours (range, 0–100 h) and the average month of experience was 38 months (range, 12–61 months).

All players participated in recording in a male group or a female group of two or three members, to avoid the influence of different attitudes toward the opposite gender on our analysis. Each group member was asked to participate in game events together as a party to keep talking with each other. They were also asked to use a voice chat system and were prohibited from using a text chat function provided by the online game system when they talked with each other.

The MMORPG used for the recording were Ragnarok Online, Monster Hunter Frontier, and Red Stone, which each player actually played and enjoyed in their daily life. The most popular game was Ragnarok Online (7 players).

2.3. Recording environment

Figure 1 shows our recording environment. Each player played the online game at a remote location. The players wore headset microphones, AudioTechnica Dynamic Headset ATH-30COM, and chatted with each other in non-face-to-face situations via the voice chat system, Skype [4]. The voice-recording system for Skype, Tapur [5], was used to record their chats. Tapur recorded the local player's voice and remote player's voices in individual channels of a stereo sound file. As the local player's voice was recorded directly into the recording PC, there was no distortion or encoding effects in local sound data caused by transmission via the Internet. The recording time was 1 hour for each group. The total recording time was approximately 13 hours. The sound data were sampled at 48 kHz and digitized to 16 bit. The database includes only the local player's voice.

2.4. Segmentation

Recorded dialogs were segmented into utterances. Any continuous speech segment between pauses exceeding 400 ms was regarded as a unit of utterance. As a result, the total number of utterances in our database was 9,114. Some utterances were excluded from the target utterances for annotation of perceived emotion, because these utterances were inadequate for acoustic analysis. As a result, the number of the total utterances for annotation was 6,578 utterances.

3. Annotation of perceived emotion

3.1. Method

To specify one perceived emotion to each utterance, the

*e-mail: ar@mf.teu.ac.jp
†Current affiliation: JST-ERATO Okanoya Emotional Information Project

Table 1 Confusion matrix between majority-agreed emotions vs the remaining annotator's choice.

		remaining annotator's choice									
		FEA	SUR	SAD	DIS	ANG	ANT	JOY	ACC	NEU	OTH
majority-agreed emotion	FEA	**23%**	17%	15%	11%	2%	6%	4%	1%	11%	10%
	SUR	6%	**31%**	3%	6%	7%	11%	12%	6%	12%	7%
	SAD	13%	5%	**20%**	14%	4%	4%	5%	9%	15%	11%
	DIS	7%	4%	12%	**13%**	19%	3%	4%	9%	22%	6%
	ANG	3%	9%	3%	29%	**25%**	7%	5%	6%	11%	3%
	ANT	3%	5%	3%	6%	4%	**16%**	22%	9%	23%	11%
	JOY	2%	9%	5%	4%	3%	18%	**29%**	13%	12%	5%
	ACC	3%	3%	7%	11%	2%	11%	12%	**9%**	35%	7%
	NEU	5%	6%	7%	10%	5%	14%	7%	22%	**15%**	9%
	OTH	6%	15%	13%	6%	4%	6%	6%	7%	24%	**15%**

The highest rate of confusion of each agreed emotion with the other emotions except NEU and OTH, is highlighted in gray. The rate of confusion of each agreed emotion with NEU is also highlighted in gray.

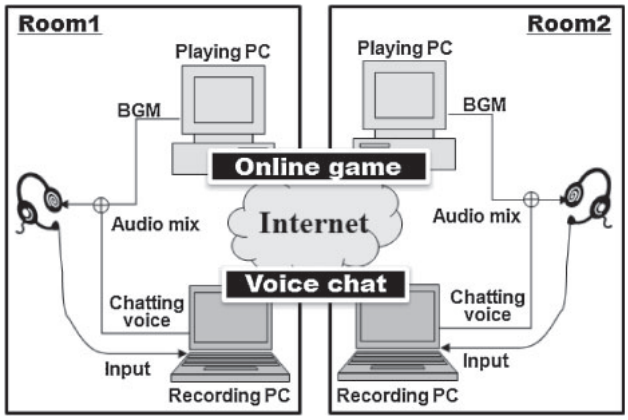

Fig. 1 Recording environment.

annotation of perceived emotion was conducted. Annotators were forced to choose one emotion for each utterance from 10 alternatives of 8 emotions (fear (FEA), surprise (SUR), sadness (SAD), disgust (DIS), anger (ANG), anticipation (ANT), joy (JOY), and acceptance (ACC)) and neutral (NEU), which does not include any emotion, and emotions other than the above 8 emotions (OTH). Eight emotions were prepared with reference to the primary emotions in Plutchik's multi-dimensional model [6]. The explanation of all 10 alternatives for annotation were prepared with reference to a dictionary [7] and presented to the annotators in order to fill gaps between the definitions of emotions among annotators.

Twenty-two annotators shared the load of listening and judging 6,578 utterances. Every utterance was judged by three different annotators. Each utterance was presented once to each annotator in random order to avoid any order effect. The annotators were requested to judge utterances on the basis of those acoustic characteristics, not the content of the utterance.

3.2. Results

The perceived emotion for each utterance was approved by a majority vote. The agreement rate was 58%, which was calculated by dividing the number of utterances on which a majority of annotators (two out of three) agreed, by the total number of utterances (6,578 utterances). Table 1 shows the confusion matrix of majority-agreed emotions, where two out of the three annotators were in agreement, and the emotions chosen by the remaining annotator. The majority-agreed emotions are shown in the rows and the remaining annotator's choice is shown in the columns of the table. The agreement rates among all three annotators are aligned diagonally and shown in bold face.

SUR, JOY, ANG and FEA show the highest agreement rates among all three annotators (31%, 29%, 25%, and 23%). The rate of confusion of each majority-agreed emotion with NEU ranges from 11% to 35%. The highest rate of confusion of each majority-agreed emotion with the other emotions, except NEU and OTH, ranges from 11% to 29%.

4. Rating emotional intensity

4.1. Method

A listening test was conducted to quantify the perceived intensity of emotion of each utterance of the 8 emotions. The evaluators were 18 university students (13 males and 5 females). Each utterance was grouped into each of the majority-agreed emotions, and was presented once in random order to avoid any order effect. The evaluators were informed of the kinds of emotion of utterances, and were directed to rate each utterance on a five-point scale from 1 (weak emotion) to 5 (strong emotion).

The NEU and OTH utterances were excluded from the dataset for the listening test for the following reasons. The NEU utterances were regarded by the annotators as those of no emotion as a result of the annotation described in Section 3. The NEU utterances were excluded from the dataset because it was impossible to judge the intensity of emotion for utterances of no emotion. The OTH utterances were regarded as those including one or more emotions, other than Plutchik's 8 primary emotions, but the kind of emotion for each utterance was not specified by the annotators. In the listening test, the kinds of emotion of utterances were informed to the evaluators. However, it was impossible to present the kind of emotion of the OTH utterances because it was not specified. The OTH utterances were also excluded from the dataset because it is confusing for the evaluators to judge the intensity of an unspecified emotion for each OTH utterance, and hence, it was considered that reliable judgments could not be obtained in the listening test.

Although, the evaluators were requested to rate the intensity of the emotion of utterances from the acoustic characteristics, not from the linguistic content of the utterance, three of the evaluators reported that they judged the emotional intensity of utterances from the linguistic content of the utterances. Therefore, the scores assigned by these three evaluators were removed from the following analysis.

4.2. Results

The interevaluator agreement rate was calculated using Spearman's rank correlations. Average correlation coefficients were 0.29 for the 8 emotions as a whole, 0.25 for FEA, 0.38 for SUR, 0.17 for SAD, 0.17 for DIS, 0.21 for ANG, 0.20 for ANT, 0.33 for JOY, and 0.29 for ACC. The rated emotional intensity of each utterance is considered to spread at random in intensity dimension of psychological emotional space because no directions were given to the speakers to act any emotion at a certain level of emotional intensity and they were only requested to play the game as they usually did. To obtain continuous emotional intensity at irregular intervals from discrete scores at even intervals, the mean value of all scores judged by the evaluators was calculated as the emotional intensity for each utterance.

5. Discussion

Two assumptions are made about the nature of confusion in perceived emotions and the results of annotation and rating are discussed from the viewpoint of intensity and similarity dimensions. The first assumption about the nature of confusion is derived from one emotion of weak intensity and the second assumption is derived from the similarity of emotions.

The first assumption is derived from the phenomenon that confusing utterances with NEU are often observed in our annotation. As NEU does not include any emotion, a mixed emotion of any emotion and NEU is considered to be a weak intensity of emotion. Figure 2 supports this assumption with showing the distributions of emotional intensity vs utterances of emotion agreed upon by all three annotators (complete-agreement group) and utterances on which two annotators were in agreement and one annotator chose NEU (partial-agreement group). The partial-agreement group shows a lower mean value of intensity than does the complete-agreement group for all emotions except SAD. Student's t-test was also conducted to verify the null hypothesis that there is no difference between the mean values of the two groups. As a result, the mean values of the two groups for all emotion, except SAD, were significantly different ($\alpha < 0.05$). Thus, confusing utterances with NEU had weak intensity of emotion.

The second assumption is derived from the highest rate of confusion of each majority-agreed emotion with an emotion other than NEU and OTH. The emotion most often confused with the majority-agreed emotion is considered to be the similar emotion to the majority-agreed emotion itself. The rates of confusion of the majority-agreed emotions with emotions chosen by the annotator, are listed in Table 1. For example, the emotion most often confused with the majority-agreed FEA is SUR, and its rate of confusion is 17%, and the least often confused is ACC (1%). According to Plutchik [6],

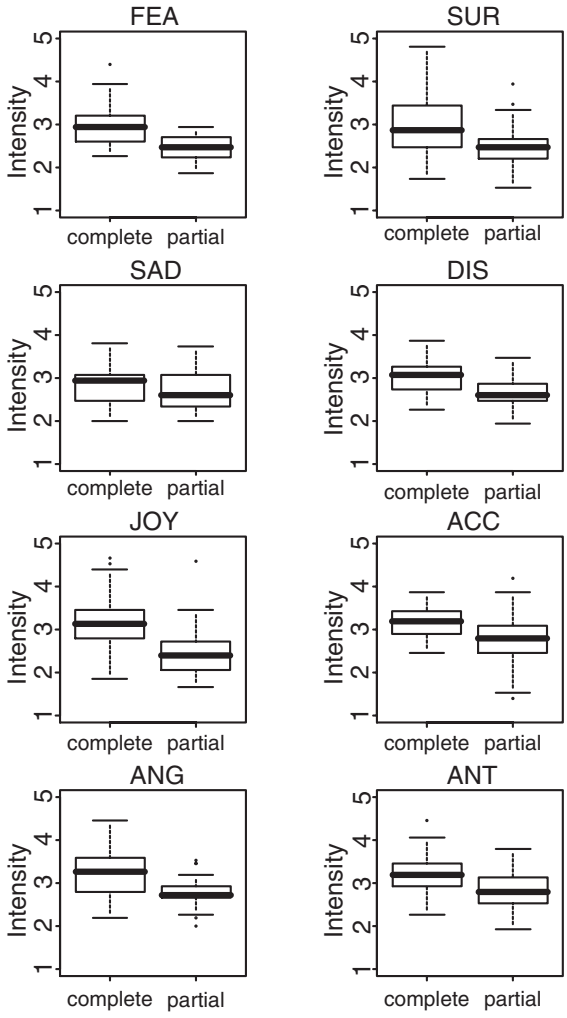

Fig. 2 Distributions of intensity vs utterances of emotion agreed on by all three annotators (complete), and utterances on which two annotators agreed and one annotator chose NEU (partial).

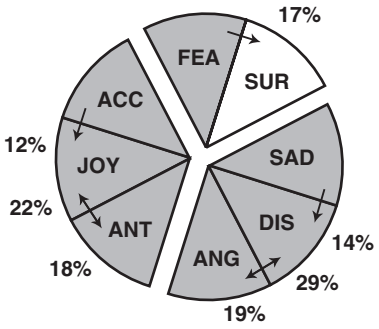

Fig. 3 Diagram of the rates of confusion and Plutchik's emotion wheel from the similarity perspective.

adjacent pairs on the emotion wheel of 8 primary emotions reflect emotional similarity; for instance, joy is placed between acceptance and anticipation so that joy is similar to both acceptance and anticipation. Figure 3 is a diagram showing the rate of confusion in Table 1 on an emotion wheel

with reference to the emotion wheel of Plutchik's 8 primary emotions. Sectors on the wheel in Fig. 3 are placed apart when adjacent emotions do not indicate the highest rate of confusion with each other. The sectors of emotions that show the highest rate of confusion with the adjacent emotion are shaded gray. The direction of arrows and the numbers indicate the confusion direction and the rate of confusion. For instance, FEA, shaded gray, is often confused with SUR at the rate of confusion of 17%. Seven out of 8 emotions show the highest rate of confusion with one of the adjacent emotions. Figure 3 indicates that most of the emotions show the highest rate of confusion with the adjacent emotions. This result suggests that these emotions were perceived to be similar to the adjacent emotions.

SUR is the only exception to the second assumption. The emotion most often confused with majority-agreed SUR is JOY, which is not an adjacent emotion of SUR. Its rate of confusion is 12%. Ekman [8] pointed out that surprise is the briefest of all emotions, lasting only a few seconds at most, and surprise passes quickly as we figure out what is happening. Then, surprise merges into other emotions such as fear, anger, and so forth, depending on what it was that surprised us, or it may be followed by no emotion at all if we determine that the surprising event is of no consequence. Our spontaneous emotional speech database comprised utterances from dialogs among online game players chatting via a voice chat system. Most of the game players participated in the recording because of their desire to enjoy an online game with their partners. Owing to the brief-lasting and variable nature of surprise and the characteristic of our database, surprise of the game players quickly merged into the joy of playing the game even during one utterance. The annotators perceived both surprise and joy in one utterance, and it lead to the high rate of confusion of SUR with JOY.

6. Conclusion

In this letter, we assessed our spontaneous emotional speech database comprising utterances from dialogs among online game players chatting via a voice chat system, and discussed the perceived emotions from the viewpoint of intensity and similarity.

As a result, we found that the intensity and similarity of emotion are two of the components of the emotional space perceived by humans when they listen to emotional speech. In future work, we will investigate correlations between acoustic features and emotional intensity in order to the realize automatic emotional intensity estimation.

Acknowledgment

The authors express their gratitude to Dr. Hiromi Kawatsu for her contribution to designing the database.

References

[1] The National Institute for Japanese Language, *Construction of the Corpus of Spontaneous Japanese* (The National Institute for Japanese Language, Tokyo), (2006) (in Japanese).
[2] H. Mori, H. Aizawa and H. Kasuya, "Consistency and agreement of paralinguistic information annotation for conversational speech," *J. Acoust. Soc. Jpn.* (*J*), **61**, 690–697 (2005) (in Japanese).
[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, **18**, 32–80 (2001).
[4] Skype, http://www.skype.com/intl/ja/.
[5] Tapur, http://www.tapur.com/jp/.
[6] R. Plutchik, *Emotion — A Psycho-evolutionary Synthesis* (Harper & Row, New York, 1980).
[7] T. Yamada, T. Shibata, Y. Kuramochi and A. Yamada, *Shin-Meikai Japanese Dictionary*, 6th ed. (Sanseido, Tokyo, 2005) (in Japanese).
[8] P. Ekman, *Emotions Revealed* (Owl Books, New York, 2003).