interspeech 2008 incorporating SST 08

Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data

Khiet P. Truong[1], *Mark A. Neerincx*[1,2] *and David A. van Leeuwen* [1]

[1]TNO Defence, Security and Safety, P.O. Box 23, 3769ZG, Soesterberg, The Netherlands
[2]Delft University of Technology, P.O. Box 5031, 2628CD, Delft, The Netherlands
{khiet.truong,mark.neerincx,david.vanleeuwen}@tno.nl

Abstract

We investigated inter-observer agreement and the reliability of self-reported emotion ratings (i.e., self-raters judging their *own* emotions) in spontaneous multimodal emotion data. During a multiplayer video game, vocal and facial expressions were recorded (including the game content itself) and were annotated by the players themselves on arousal and valence scales. In a perception experiment, observers rated a small part of the data that was provided in 4 conditions: audio only, visual only, audiovisual and audiovisual plus context. Inter-observer agreements varied between 0.32 and 0.52 when the ratings were scaled. Providing multimodal information usually increased agreement. Finally, we found that the averaged agreement between the self-rater and the observers was somewhat lower than the inter-observer agreement.

Index Terms: emotion, multimodal database, inter-rater agreement, self-reported emotion

10.21437/Interspeech.2008-95

1. Introduction

In affective computing, there is a growing need for automatic emotion analyzers that can analyze complex, spontaneous emotions which are expressed via multimodal channels in real-world environments. In order to develop these emerging emotion analyzers, multimodal spontaneous, annotated emotion databases need to be developed. One of the difficulties in acquiring such data is that emotion is a highly subjective phenomenon for which it is difficult to define a 'ground truth' that is required for an algorithm to 'learn' to recognize emotion. Usually, an annotation can be considered reliable when multiple raters, who have annotated the same data, agree with each other. The reliability of spontaneous (multimodal) emotion data has been investigated previously by e.g., [1, 2, 3, 4, 5]. In a first trial, Reidsma et al. [2] obtained low averaged pair-wise agreements between 0.07 and 0.18 on real-life multimodal meeting data. Laskowski and Burger [1] report inter-rater κ agreements between 0.15 and 0.67 for three annotators who performed valence annotation on meeting data. Douglas-Cowie et al. [5] report inter-rater κ agreements between 0.37 and 0.54 where two annotators annotated spontaneous multimodal emotion data in a category-based approach. They found that agreement was lowest when multimodal (audiovisual) emotion data extracted from television was provided to the raters.

However, in the context of automatic emotion recognition, the reliability of annotations performed by people annotating their *own* emotions have not been investigated yet. A high agreement between observers and the self-rater indicate that self-ratings are reliable and that 'felt' emotion can be perceived. We propose to investigate the reliability of self-reported emotion ratings (i.e., ratings of people who have rated their own emotions) and make the assumption that these self-reported emotion ratings lie closest to the 'ground truth' emotion. To that end, we recorded a novel multimodal spontaneous emotion database. Since gaming can provide a natural though controlled setting in which a participant can immerse and can express his/her emotions, we decided to elicit emotions via a multiplayer first-person shooter video game.

In this study, we use this database to investigate human recogniton of multimodal spontaneous emotion. In this paper, we aim to assess 1) how well observers agree on the perception of spontaneous emotion (i.e., inter-observer agreement), 2) how well observers agree when audio only, visual only, audiovisual or audiovisual plus context information is provided, and 3) the reliability of self-reported emotion ratings.

In Section 2, we describe how the data was collected and annotated by the participants themselves. In Section 3, we explain the setup of the perception experiment and how agreement is calculated. The results of the agreement analysis among the observers are presented in Section 4, as well as the results of the reliability analysis of the self-ratings. Finally, we summarize and discuss our conclusions in Section 5 and elaborate on future research.

2. Acquiring spontaneous multimodal emotion data and self-reported emotion ratings

2.1. Participants

Seventeen males and eleven females with an average age of 22.1 years (2.8 standard deviation) participated in an experiment by playing a multiplayer video game. We asked each participant to bring along a friend as his /her team mate. A compensation was paid to all participants and bonusses were granted to the winning team, and the team with 'best collaboration'. The latter bonus was initiated to encourage the participants to be vocally expressive.

2.2. Recordings

Speech recordings were made with high quality close-talk microphones that were attached near the mouth to minimize the effect of crosstalk (overlapping speech from other speakers) and other background noise. Recordings of facial expressions were made with high quality webcams. The webcams were placed at approximate eye-level on top of the monitor such that a frontal view of the face was recorded under an angle that was acceptable for reliable automatic facial recognition software. Further, lighting and background conditions were controlled by adjust-

Accepted after peer review of full paper

September 22 – 26, Brisbane Australia

ing the light whenever needed and by placing evenly coloured dark curtains behind the participants to avoid clutter in the background. The game content itself was also stored by capturing the frames of the video stream during game play. In Fig. 1, some examples of recordings are shown.

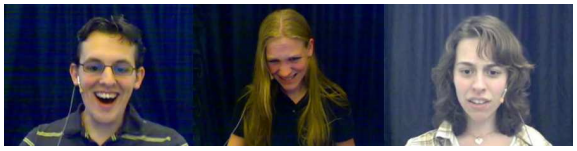


Figure 1: *Examples of facial expressions in the database.*

2.3. Procedure

In teams of 2×2, the participants played a first-person shooter video game called *Unreal Tournament* by Epic Games. We selected the game mode 'Capture the flag' and a very small 3D world in which the goal was to capture each other teams' flag as many times as possible. During the game, at an approximate rate of one event per minute, the experimenter generated 'surprising' game events with the game engine in order to emotionally arouse the players. Sudden deaths, the sudden appearance of a bunch or monsters, hampering keyboard or mouse controls etc., were some of the events that were generated. There were two game sessions, each of 20 minutes long. Prior to each game session, the participants performed a training session to get acquainted with the game (10 minutes) and received instructions and a training session (40 minutes) involving the emotion rating tasks. After each game session, the participants performed the emotion rating tasks (50 minutes).

2.4. Self-reported emotion ratings

Each participant gave his/her own emotion ratings based on the 3 types of information that were registered during the previous game session: the audiovisual content containing (1) vocal and (2) facial expressions in frontal view, and the video of the (3) game content that was captured. The participants were asked to recall what emotions they experienced at that moment in the game while listening and watching the audiovisual contents. The videos were shown continuously and the participants could not pause the streams. They performed the two following annotation tasks based on two approaches: 1) discrete category-based (i.e., make a forced choice between a number of emotion labels), and 2) (semi-)continuous dimensional-based approach. In this paper, we will focus on the dimensional- based approach.

In the dimensional-based rating task, we assume that emotions can be described by two dimensions namely arousal (active vs. passive) and valence (positive vs. negative). The third dimension 'dominance' will not be used here. We employed a semi-continuous approach for the dimensional-based task; each 10 seconds, participants had to give arousal and valence ratings separately (thus *not* simultaneously as is the case in Feeltrace), on a scale from 0 to 100 with 50 being neutral.

The results of the self-reported dimensional-based approach is presented in Fig. 2. A striking similarity between our arousal-valence plot based on self-reported ratings and that of Bradley&Lang [6] appears. Although our context differs from that of Bradley&Lang [6] (listening to acoustic stimuli), we can reproduce a similar 'boomerang' shape-like distribution of arousal-valence ratings. We found a quadratic relationship between arousal and valence: high or low valence often co-occurs with high arousal.

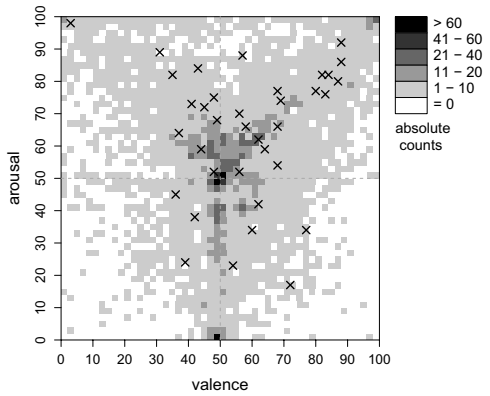


Figure 2: *2D histogram plot: self-reported (raw) arousal and valence ratings (N=6870, nbins=50). '×' represents the means of the 36 video clips that were presented to observers in the perception experiment.*

3. Assessing inter-observer agreement in a perception experiment

A perception experiment was carried out to assess inter-observer agreement. A selection of movie clips was rated by observers in the same way as the self-rating. The movie clips originated from various regions in the arousal-valence space and were offered to the observers in various conditions. In this Section, we describe the setup of the perception experiment and how agreement was calculated.

3.1. Setup perception experiment

We invited twelve female and six male observers with an average age of 21.9 years (we will also use the term 'raters' to refer to these observers) to rate a small part of the spontaneous audiovisual emotion data in a similar way as the players themselves had done. All 18 observers had not participated in the game sessions. We selected six different players, each from which six movie clips were selected by a number of criteria: the movie clips had to contain a sufficient amount of vocal *and* facial expressions, and the aim was to have movie clips originating from different regions in the arousal-valence space. These regions include the four well-known quadrants and 2 'emotion changes': positive-active (PA), positive-passive (PP), negative-active (NA), negative-passive (NP), large change in arousal (CA) and a large change in valence (CV). This makes a total of 6×6 movie clips that were presented to each observer. However, it appeared to be difficult to satisfy all of these criteria. As a result, not all emotion quadrants were equally well represented in the set of stimuli (see Fig. 2). Each movie clip had a length of 55 seconds and 4 rating moments. At each rating moment, an arousal and valence rating had to be given, similar to the rating task that was performed by the players themselves.

The movie clips were presented to the observer in six different conditions: audio only (A), visual only (V), audio+context (AC), visual+context (VC), audio+visual (AV), audio+visual (AV) and audio+visal context (AVC). With 'context' we mean the game content that was recorded during game play. The AVC

condition is best comparable to the players' rating task in which self-ratings were performed with audio plus visual plus context information.

In a within-subject design, the 36 movie clips were distributed over 36 cells in a 6 (conditions) $\times$ 6 ('emotion regions') matrix and presented to the observers in a balanced design such that each movie clip of a specific player with a specific 'emotion region' was rated in each condition by at least two observers, see Table 1 for one example design for one observer.

Table 1: *Example of distribution of movie clips over conditions and 'emotion regions' for one observer.*

	'emotion region'					
cond	PA	NA	CA	CV	PP	NP
A	player1	player2	player3	player4	player5	player6
V	player1	player5	player4	player2	player3	player6
AV	player4	player2	player3	player1	player5	player6
AC	player2	player6	player3	player5	player1	player4
VC	player1	player6	player2	player3	player5	player4
AVC	player3	player6	player5	player2	player4	player1

3.2. Computing agreement

We used Krippendorff's Alpha α [7] (ordinal scale) to assess the agreement between multiple raters. For each emotion dimension, there were 144 possible rating moments (36 movie clips with each 4 ratings). We chose to have a within-subjects design that is balanced but incomplete in the sense that not all movie clips were rated by all same observers. Each movie clip is rated by at least 2 observers. In assessing the reliability of content data where multiple raters are used to annotate the data, it not uncommon that raters code different subsamples of the data. According to Krippendorff, "coders must be interchangeable, may code different subsamples of units of analysis, provided there is enough duplication or overlap". Krippendorff's α is flexible enough to accommodate for 'missing values'. α has a range from $[-1, 1]$. When observers agree perfectly, α is 1. When α is 0, agreement is by chance and when α is smaller than 0, it indicates disagreement. Prior to calculating α, the ratings were discretized into 3 (boundaries on 35 and 65) or 5 (boundaries on 20, 40, 60 and 80) classes.

Furthermore, in addition to the use of raw emotion rating values for the calculation of agreement, we also computed the *changes* between subsequent emotion ratings to evaluate whether people judge emotion better in a *relative* manner than an *absolute* manner. These *delta* ratings were computed by subtracting the previous rating from the current rating in each movie clip, see Fig. 3. Finally, to adjust for personal differences between observers, i.e., some observers tend to use the whole scale, while others only use a small part of the scale, we linearly scaled all the arousal and valence ratings to a range of $[0, 1]$ per observer.

4. Results

In this section, we present the results of the perception experiment. The movie clips were presented to the observers under various conditions; we report results obtained in the A, V, AV and AVC conditions. Furthermore, since there were minimal differences between the agreement figures that were based on discretization in either 3 or 5 classes, we report agreement results that are based on a discretization in 5 classes. Finally,

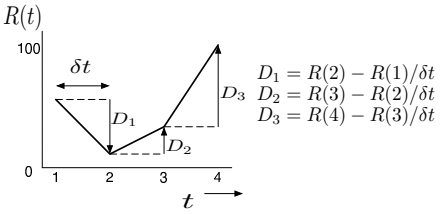
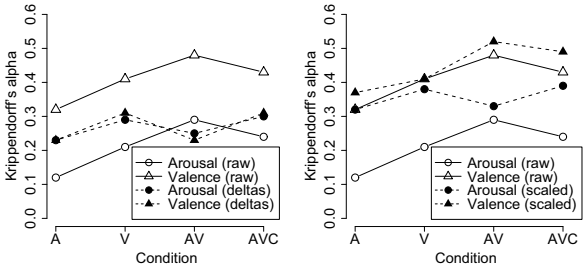


Figure 3: *Computation of 'delta' ratings for each movie clip, in this case δt is 1, because each rating $R(t)$ is given at a fixed interval δt ($R(t)$ is an emotion rating given at moment t).*

we compare the observers' ratings to the ratings of the players themselves.

4.1. Inter-observer agreement: agreement between multiple observers

The results are presented in Fig. 4. We can observe in Fig. 4(a) that inter-observer agreement based on raw ratings ranges from 0.12 to 0.48: the highest αs are observed in the AV condition. Apparently, observers do benefit from the multimodal information that is made available to them, although the addition of context does not seem to offer additional information. The visual channel seems to provide more information than the acoustic channel. Furthermore, inter-observer agreement is sytematically worse for arousal than valence. However, when we use the 'delta' ratings, α increases considerably for arousal, but not for valence (see Fig. 4(a)): this suggests that people are better able to judge *changes* of arousal rather than *absolute* arousal.



(a) Raw and Deltas (b) Raw and Scaled

Figure 4: *Krippendorff's α inter-observer agreement.*

In Fig. 4(b), we can observe that linearly scaling the ratings to $[0, 1]$ leads to a substantial improvement of α for arousal. For valence, this improvement is small. α now ranges from 0.32 to 0.52 which is higher than α based on raw ratings: this improvement is largely due to the increase of α on the arousal dimension. Furthermore, similar to the raw case, arousal is worse agreed upon than valence, multimodal information is beneficial, and visual information is stronger than acoustic information.

4.2. Agreement between the self-rater and multiple other observers

We analysed agreement between the self-rater and the observers by adding the ratings of the self-rater to the group of 18 observers. If α does not decrease, it indicates that the added self-rater did not influence the inter- observer agreement negatively.

This might imply that the observers agree with the self-rater, and that observers have the ability to perceive 'felt' emotion. However, in Table 2, we can observe that the addition of the self-rater affects α negatively.

Table 2: *Krippendorff's α inter-rater agreement, either without or with the self-rater (+self), for the AVC condition.*

	raw		deltas		scaled	
		+self		+self		+self
arousal	0.24	0.23	0.30	0.21	0.39	0.36
valence	0.43	0.37	0.31	0.29	0.49	0.40

An alternative way to assess the reliability of self-ratings is to compute pair-wise agreements between an observer and the self-rater, which also enables assessment of individual performances. The averaged, minimum and maximum αs are shown in Table 3. The highest averaged α s are 0.30 and 0.34 for arousal and valence respectively, which are somewhat lower than the inter-observer αs observed in Fig. 4(b) and Table 2 in the AVC condition. Furthermore, the minimum and maximum αs indicate that there are large individual differences between the observers (Table 2).

Table 3: *Krippendorff's α for pair-wise agreement between a rater and the self-rater in the AVC condition.*

	arousal			valence		
	mean	min	max	mean	min	max
raw	0.16	-0.27	0.51	-0.09	-0.45	0.34
deltas	0.13	-0.37	0.48	0.24	-0.29	0.69
scaled	0.34	-0.07	0.70	0.30	-0.21	0.62

5. Discussion and conclusions

To summarize, we have investigated inter-observer agreement among multiple observers that have rated multimodal spontaneous emotion data on arousal and valence scales. With αs ranging from $0.32 - 0.52$ when the data is scaled, we achieved inter-observer agreement figures which are in line with previous results in [1, 2, 5]. Scaling appeared to be necessary for adjusting for the fact that raters have different emotion ranges. Scaling all ratings (linearly) to a scale of $[0, 1]$ improved agreement considerably.

When we compare the agreement figures per emotion scale, we find that agreement is usually higher on the valence scale. Improvements for arousal can be achieved when the *relative changes* in arousal are taken into account instead of the absolute values. This does not seem to apply for the valence scale.

Finally, we found that using multimodal instead of unimodal information usually improved agreement. Using AV or AVC information, instead of A or V only, improved agreement in the majority of cases when the ratings were raw or scaled; this improvement was more apparent for the valence than the arousal scale. In a previous study, Douglas-Cowie et al. [5] found that AV seemed to complicate and confuse emotion recognition in a categorical-based rating task (but note that their study was based on naturalistic emotion data extracted from television). We also found that in most of the cases, the V condition led to higher agreement than the A condition. Furthermore, the addition of context information did not result in consistent results. One of the reasons why this effect is not apparent might be that context information can be provided in various manners, in our case, a visual manner. The game content was shown next to the visual channel with the consequence that the observer had to divide his/her attention between two screens. In addition, the movie clips provided to the observers were cut from their 'contextual flow', so that it may have been difficult to really understand the context.

We have also investigated the reliability of self-reported emotion ratings. We found indications that self-ratings differ, in some cases substantially, from the observers' ratings. However, the averaged agreement between the self-rater and the observers lie between 0.30 and 0.34 and are in comparison to the inter-observer agreement figures not bad. Further investigation regarding the use of self-reported emotions can provide more insight, when, for example, intra-reliability of the self-raters and the observers can be assessed. This was not possible in the currenty study since all observers and self-raters rated the data only once.

In future research, our plan is to develop multimodal emotion analyzers based on this database, taking into account the findings of the current study. For example, the agreement achieved on the arousal scale suggests that it can be interesting to train models to detect *changes* in arousal rather than *absolute* arousal. Moreover, we will compare machine recognition to human recognition of emotion. Besides the dimensional-based emotion ratings, we will also assess agreement based on a categorical emotion rating task. Finally, we plan to investigate whether the 'emotion triggers' used in our study, i.e., the sudden game events, can be related or even predict certain emotional states as indicated by the players.

6. Acknowledgements

We would like to thank Dennis Reidsma and Paul Merkx for their support. This research was supported by MultimediaN, a Dutch BSIK project.

7. References

[1] Laskowski, K., Burger, S. "Annotation and analysis of emotionally relevant behavior in the ISL Meeting Corpus", Proceedings of LREC, 2006.

[2] Reidsma, D., Heylen, D., Ordelman, R. "Annotating emotion in meetings", Proceedings of LREC, 2006.

[3] Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macias-Guarasa, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C., Wrede, B. "The ICSI Meeting Project: Resources and Research", NIST ICASSP 2004 Meeting Recognition Workshop, 2004.

[4] Wrede, B., Shriberg, E. "Spotting 'hot spots' in meetings: human judgments and prosodic cues", Proceedings Eurospeech, 2805–2808, 2003.

[5] Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Savvidou, S., Abrilian, S., Cox, C. "Multimodal databases of everyday emotion: facing up to complexity", Proceedings of Interspeech, 813–816, 2005.

[6] Bradley, M.M., Lang, P.J. "Affective reactions to acoustic stimuli", Psychophysiology 37, 204–215, 2000.

[7] Krippendorff, K. "Computing Krippendorff's Alpha-Reliability", Available online 29/03/08, "http:/www.asc.upenn.edu/usr/krippendorff/webreliability.doc".