Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus

Tanja Bänziger and Klaus R. Scherer

Chapter to be published in Scherer, K.R., Bänziger, T., & Roesch, E. (Eds.) A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing. Oxford: Oxford University Press, in press

Note: titles and text are subjected to change, do not quote verbatim

Abstract

In this chapter we outline the requirements for a systematic corpus of actor portrayals and describe the development, recording, editing, and validating of a major new corpus, the Geneva Multimodal Emotion Portrayal (GEMEP). This corpus consists of more than 7,000 audio-video emotion portrayals, representing 18 emotions (including rarely studied subtle emotions), portrayed by 10 professional actors who were coached by a professional director. The portrayals are recorded with optimal digital quality in multiple modalities, using both pseudo linguistic utterances and affect bursts. In addition, the corpus includes stimuli with systematically varied intensity levels, as well as instances of masked expressions. From the total corpus, 1,260 portrayals were selected and submitted to a first rating procedure in different modalities to establish validity in terms of inter-judge reliability and recognition accuracy. The results show that the portrayed expressions are recognized by lay judges with an accuracy level that, in the case of all emotions, largely exceeded chance and that compares very favorably with published tests of emotion recognition that use highly selected stimulus sets. The portrayals also reach very satisfactory levels of inter-rater reliability for category judgments and ratings of believability and intensity of the portrayals.

The validity of the corpus is further confirmed by replicating results in earlier work on the role of expression modality and the corresponding communication channel for cue utilization in emotion recognition. We show that, as expected, the highest accuracy results if both auditory and visual information (voice, face, and gestures) is available, but that sizeable accuracy is achieved even when only one modality is available. The video modality is slightly superior to the audio modality, probably reflecting the fact that facial and gestural cues are more discrete and iconic than vocal cues. However, there are important interactions between emotion and modality, as particular emotions seem to be preferentially communicated by visual or audio cues. The results also raise important issues concerning the relationships

between intensity of expression, accuracy, and believability, thereby challenging earlier assumptions.

Introduction: Designing the corpus

Research on emotional expression (EE) is in large part based on facial, vocal, or bodily portrayals of particular emotions by professional or trained actors that have been photographed or audio- and/or video-recorded for systematic stimulus presentation purposes (Banse & Scherer, 1996; Ekman & Friesen, 1976; Lundqvist, Flykt, & Öhman, 1998; Goeleven, De Raedt, Leyman, & Verschuere, 2008; Gosselin, Kirouac, & Doré, 1995; Hawk, Van Kleef, Fischer, & Van der Schalk, 2009; Maurage, Joassin, Philippot, & Campanella, 2007; Johnson, Emde, Scherer, & Klinnert, 1986). The widespread use of emotion portrayals is due to three main factors: 1) Intense emotions are relatively rare, of short duration, strongly subject to social control or self-regulation, and generally occurring in an unpredictable fashion and outside of public scrutiny (see CHAPTER 3.2). As a consequence, systematic recordings of genuine, spontaneous emotions are difficult to obtain (see CHAPTER 6.2); 2) Much of emotional expression research is concerned with the communication of emotional meaning, in particular the perception of expressions by an observer and the inferences drawn from these. In consequence, the emphasis is on fairly prototypical expressions with an established signal value rather than on idiosyncratic forms of emotion externalization which may vary greatly according to the respective social context; and 3) Experimental research in psychology, the neurosciences, and affective computing needs to create a sufficient number of conditions, repetitions, and controls, as well as a high degree of standardization, which can only be obtained by using acted portrayals (see also Bänziger & Scherer, 2007). In this chapter we present a new corpus constructed according to a number of theoretical and methodological desiderata and based on our past experience with similar efforts. Because this corpus of actor portrayals was planned to be shared with different research communities for different purposes, careful planning of the set of emotions to be included and the type of portrayal instructions was needed.

Selection of emotions portrayed. The affective states selected for portrayal were partly chosen to represent the states that are frequently studied in the literature that deal with facial and/or vocal expressions of affect, in particular the set of basic emotions as defined by basic emotion theorists (Ekman, 1999, Izard, 1991). Less frequently examined states were also included to address specific research questions. For example, a relatively large number of positive states – such as pride, amusement, elation, interest, pleasure, and relief – were included to challenge the traditional view in which only one undifferentiated positive state (happiness) can be reliably communicated via facial cues. In a similar attempt, some states corresponding to the same family of emotional reactions were included with various arousal levels (e.g., irritated and enraged anger; anxious and panic fear). This inclusion fulfills two aims:

1. Reviews of studies describing acoustic profiles of EEs have repeatedly reported differences in acoustic features of vocal expressions mostly related to arousal level. Variations of arousal within emotion "families" should allow us to disentangle the influence of arousal level and emotion family on vocal expressions.
2. The inclusion of more than one type of anger (or fear) should result in increased variability of the expressions portrayed and should allow us to include a range of variations that are more likely to occur in daily interactions, under the assumption, for example, that irritation occurs more frequently than rage and anxiety more frequently than panic fear.

In choosing the set of emotions, an attempt was made to theoretically order them for a valence and an arousal dimension. This is shown in Table 1, together with a list of additional emotions added for the reasons outlined earlier.

Table 1 about here

A further attempt to increase the variability of the expressions was undertaken by including portrayals for some of the emotions with less and with more emotional intensity than what corresponds to the usual intensity for a given emotion (baseline intensity). An underlying assumption is that the portrayals produced with less intensity might be closer to expressions that could occur in daily interactions, whereas the portrayals with more intensity might be more exaggerated (or more stereotypical). To this regulation of the intensity of portrayed states, we added a further condition to partially mask some of the expressions (i.e., to portray an unsuccessful deception attempt for some of the affective states).

The role of different expression modalities in the recognition of emotion. Much of the work in using actor portrayals has been interested in cue utilization in the recognition of the intended emotion by observers, an essential aspect of the study of pull effects. Much of this literature has been concerned with the role of different expression modalities and the respective communication channels, for example, vocal expression as carried in the audio modality versus facial and gestural expression as carried in the video modality. Because traditionally researchers have specialized in the domain of the face (the large majority) or the voice (a small minority), most corpora have been recorded in only one modality. In the case of facial expression, much of the work has even been limited to static photographs (which essentially eliminates the important dynamic character of unfolding emotion). As in our earlier actor portrayal studies (Banse & Scherer, 1996; Bänziger, Grandjean, & Scherer, 2009; Scherer & Ellgring, 2007), we feel that it is essential to work with multimodal corpora in which the portrayed expressions of each actor are recorded in both the video and audio mode. Only such multimodal corpora allow us to study the relative importance of audio and video cues both in isolation and in interaction. The GEMEP corpus described in the following sections is such a multimodal stimulus set, in which high-quality video recordings (including different close-ups of upper body and face, frontal and side view) and digital audio recordings have been obtained.

This chapter, apart from describing the procedures used to produce the corpus and the results of the validation studies, also presents first data on the effects of available modalities on the accuracy of emotion recognition and the perceived intensity and believability of the portrayals. From the patterns of results reported in the literature (Scherer, 1999), we hypothesized that the availability of both auditive (voice) and visual (face, gestures, posture) cues should lead to maximal accuracy. However, we also expected emotion by modality interactions; that is, some emotions being better recognized in the audio-only condition and others in the video-only condition. We also examine the role of the type of utterance used (two different nonlinguistic sentences and the sustained vowel “aaa”) on the accuracy of emotion recognition and the perceived intensity and believability of the portrayals.

Methods used to produce and validate the corpus

This section consists of two parts: (a) a detailed description of the production of the corpus (portrayal procedures, recording, and clip selection) and (b) a description of the methods used to gather data on the recognition of the portrayed emotions in the master set. We expect this description (including the Supplementary Materials[1]) to be used as the reference for all future analyses or tests derived from this corpus.

Production of the Corpus

Actors. Ten actors (5 females) were recruited with the help of a professional director (Andrea Novikov), who also supervised the acting during the recordings. The 10 actors are all professional theater actors living and working in the French-speaking part of Switzerland. They were hired for 1 day each at professional rates.

Portrayal procedures. Several weeks before coming to the laboratory, actors received the list of the emotions they were to portray, together with short definitions of the emotion and brief scenarios to illustrate the labels (see Table 2 for English translations of the original

French definitions). Three scenarios were created to instantiate each affective state (see Supplementary Materials). A scenario includes the essential features of a situation, which is assumed to elicit a given emotional reaction. Whenever possible, the scenarios included explicit references to one or more interaction partner(s). The actors were requested to improvise interactions with the director, in which they expressed a given affective state. In addition to the written instructions, the actors participated in a preparatory session with the director.

Table 2 about here

The 12 categories presented in the four cells of Table 1 were portrayed by all 10 actors. The six additional categories presented below the table were split into two groups and were portrayed by five actors each. The portrayals of the emotions allotted to a particular actor were produced, with frequent pauses, in the order chosen by the director. The portrayals were produced in an interactive setting, with the director serving as the addressee of the expressed emotions. On the basis of the structural descriptions of the emotions that had been agreed upon with the director, he interactively produced an appropriate mood in each actor by eliciting personal life events in which the actor had experienced the emotion. The procedure followed the philosophy of the Stanislavski acting method. Portrayals for each emotion were repeated until the director and the actor were satisfied with their performance.

Modulations of intensity and attempts at masking (regulations) were produced only for the states represented in the four cells of Table 1 (i.e., not for the additional states represented under Table 1). Each actor produced the three regulations (less intense, more intense, and masked) for six of the categories represented in Table 1. Different actors produced modulations of intensity and masking for different subsets of emotions. Actors 1, 3, 6, 7, and 8 portrayed regulations for hot anger (hot), despair (des), anxiety (anx), amusement (amu), interest (int), and pleasure (ple). Actors 2, 4, 5, 9, and 10 portrayed regulations for pride (pri), joy (joy), relief (rel), panic fear (pan), irritation (irr), and sadness (sad). The instructions made it clear for the actors that an emotion could have different degrees of intensity as reflected in the expressions “I felt a little anger” or “I was very angry.” The instruction to portray an emotion and to mask it simultaneously might appear paradoxical, but it made perfect sense for the actors who frequently meet this kind of request to impersonate a character who is placed in a situation that elicits strong emotions but who tries to conceal them.

Types of utterances. As the basis for their expression portrayals, the actors were asked to utter, at the apex of the relived or simulated emotion, the following two standardized sentences: (a) “ne kal ibam sud molen!” and (b) “kun se mina lod belam?” These pseudo-linguistic phoneme sequences were chosen with the help of a phonetician to represent plausible phoneme combinations, with potentially similar pronunciations in a variety of Western languages. The actors were free to imagine different types of semantic meaning while uttering the meaningless sentences. They were further requested to express each affective state while uttering a sustained vowel (“aaa”), which allowed the recording of brief EEs, reminiscent of affect bursts or interjections (Scherer, 1994), in the absence of articulatory movements.

Recording technology and procedures. The portrayals were recorded in one of the interaction laboratories of the Geneva Emotion Research Group at the University of Geneva. Three digital cameras (SONY DSR-PDX10) were used to simultaneously record: (a) facial expressions and head orientations of the actors, (b) body postures and gestures from the perspective of an interlocutor, and (c) body postures and gestures from the perspective of an observer standing to the right of the actors (see Figure 1). Sound was recorded by using three separate microphones located at each of the three cameras, plus an additional headset microphone (SENNHEISER) positioned over the left ear of the actor, providing a separate speech recording with a constant distance to the actor's mouth. The audio and video streams were recorded on four separate PCs by using the DV-AVI (PAL, 720x576) format for video and the PCM WAV (41 kHz) for audio. Each recording session lasted around 6 hr.

Figure 1 about here

Editing of individual clips. Video and audio recordings were aligned (with a precision of 1/24 s because the video cameras were not frame synchronized) and segmented on the level of single sentences. Recordings containing the two standard sentences (pseudo-speech) and the sustained vowel, as well as improvised sentences (in French), were extracted and saved into separate digital files. Over 7,300 such sequences, among them about 5,000 containing the pseudo-linguistic sentences and the sustained vowels, were extracted from the original interactions. This implies that the portrayals are extracted from ongoing interactions and therefore most often start and end with an “ongoing emotion.” This constitutes a major difference with other corpora of acted emotion portrayals, which often feature brief portrayals that start and end with a “neutral” expression.

Selection of Portrayals for a Master Set

Expert ratings were carried out to select a reduced number of portrayals with standard speech content for subsequent analyses. Three research assistants (advanced psychology students, 1 male, 2 female) were requested to assess the technical quality of the recordings and the aptitude of the actors to convey the intended emotional impression, in both vocal and facial expressions. Although the three raters showed much disagreement in their judgments, this first assessment of the portrayals allowed to observe that some actors produced a higher proportion of “convincing” portrayals than did other actors. Furthermore, there were first indications that some emotions might be more easily conveyed in either facial or vocal displays.

Based on the assessments, a selection of portrayals featuring an equal number of recordings for each actor and each portrayal category was established. Two portrayals for each condition and each actor (i.e., 126 portrayals per actor) were chosen in an iterative selection procedure by three research collaborators. Given the information provided by the expert ratings and their obvious limitations (low agreement, important rater biases, and limited number of raters), the selection had to be based on relatively complex decisions and could not be entirely systematized. In this fashion, a master set of 1,260 recordings was established.

Lay Ratings of Portrayals in the Master Set

Participants. Ninety participants, mostly undergraduate students from different departments, including psychology, were recruited via announcements in the university buildings and outside the university (e.g., in choirs of amateur singers). The participants were randomly assigned to rating either audio-only (31 participants, 18 female, 29 years on average), or video-only (31 participants, 25 female, 23 years on average), or audio-video portrayals (28 participants, 15 female, 29 years on average). The raters were paid 10 CHF for each rating session and could earn up to 100 CHF if they returned for the total of 10 sessions; however, several raters did not complete all 10 sessions. As a consequence, some portrayals were assessed by a few more raters than others (the count is provided in Supplementary Materials).

Procedure. The 1,260 selected portrayals were rated in 10 sessions of 126 portrayals produced by separate actors. A rating session always started with a set of written instructions on the rating procedure and the definitions of the emotion categories portrayed by the actors. All sessions took place in a small laboratory equipped with six computers separated by “open space” walls. Headphones were used to display the sound. One to four raters could take part simultaneously. In each session, a computer interface displayed the portrayals produced by a selected actor in two blocks: the 96 standard sentences produced by the actor were presented first in random order (a new random order for each rater was computed at the start of each session), followed by a short break and then by the 30 portrayals produced with a sustained “aaa” by the same actor, also in random order. The intensity of the sound recordings was normalized within each block to accommodate the hearing of the raters (the actors screamed in some recordings and whispered in others; the resulting variability is so large that it would not have been possible to display all recordings at a constant sound level without normalizing the sound level beforehand). The video files were compressed to a DivX format without perceptible quality loss. The video resolution was high, filling most of the screen surface. Several preset orders were defined for the successive sessions to counterbalance the sequence of actors rated. However, perfect counterbalancing was not achievable because we did not request that all raters complete the 10 rating sessions.

Presentation modalities. The ratings were collected with a computer interface, which always displayed the portrayal to be rated either in audio-only (A), in video-only (V), or in audio-video (AV) modality, depending on the randomly assigned condition for a given rater.

Instruments and procedures. First, a rating of the “believability” of each emotional portrayal was requested. Believability was rated on a continuous visual analog scale, the location of the cursor on screen being transformed to a linear scale ranging from 0 to 10. The scale was defined on screen as the “capacity of the actor to communicate a natural emotional impression” and ranged from “very low – one does not get the impression of a real emotion” to “very high – one gets the impression of a real emotion.”

Upon confirmation of the rater’s answer regarding believability, the computer displayed the 15 emotion categories portrayed by the actor on a circle (a variant of the Geneva Emotion Wheel; Scherer, 2005). The task of the participants was to select one or two categories on this circle and simultaneously rate the level of intensity (on a 4-point scale) for each of the selected categories. The emotional intensity was represented visually by the size of a bubble on screen. A legend specified that the smallest bubble corresponded to a “very weak emotion," a larger bubble to a "rather weak emotion," an even larger bubble to a "rather strong emotion," and the largest bubble to a "very strong emotion." The definitions of emotions reproduced in Table 2 were displayed on screen when the rater was moving the cursor over the respective categories (colored bubbles). The 15 categories are located on the circle according to their conceptual proximity, with positive emotions to the right side of the screen and negative emotions to the left side of the screen. Raters could select the white bubble in the center of the circle if they wished to indicate that the recording did not express an emotion. They could also click a button to type another description for the emotion portrayed in any recording (this answer was classified as "other emotion"). When a rater reported two categories, he or she had to answer a further pop-up question before proceeding to the evaluation of the next portrayal. Raters were asked to indicate if the reason for reporting two answers was either (forced choice) because those two emotions were represented in the portrayal ("mixed emotion") or because the rater was unsure and could not decide which of the two answers was "correct." The raters could replay the portrayal as often as they wished, both before rating believability and before selecting one or two categories.

Results of the validation study

In what follows, we present the results on the validity of the GEMEP Master Set in terms of 1) *reliability* (the greater the degree of agreement between judges the more reliable will be the effects of the use of the portrayals in stimulus presentations), and 2) *accuracy* (the more accurately the portrayals have been recognized the greater the likelihood that the actors produced a valid expression pattern for the respective emotions). As will be shown, the predictions on accuracy score differences between emotions and conditions vary from case to case. In the interest of readability, we do not provide details of all statistical procedures, coefficients, and exact significance levels in the text nor do we discuss peripheral or weak effects. Some of the statistical coefficients are reported in the Table notes, all other detail can be found in the Supplementary material (see Footnote 1).

Overall accuracy of emotion judgments. We computed accuracy scores for each rater in the form of a percentage of correct answers provided. An answer was defined as correct if a category reported by the participant matched the expressive intention of the actor. In cases in which a second category was reported (the instructions permitted to give one or two answers for each portrayal to allow for the perception of mixed emotions), the answer was still considered correct. Table 3 shows mean and range of the accuracy scores for raters in the three groups of raters differing in presentation modality.

The accuracy score theoretically expected by chance for 17 answer alternatives (15 emotions, no emotion, other emotion) is 5.88%. However, as two responses were allowed and given the problems of differential marginal response tendencies (see Banse & Scherer, 1996) the actual chance level is difficult to estimate. Yet, on the whole there can be no doubt that the overall accuracy levels reported in Table 3 largely exceed what could be expected by chance, which is unlikely to exceed 10-12%, providing evidence of the validity of the portrayals in terms of the encoding intentions of the actors. This is particularly the case, given the large number of emotion alternatives, largely exceeding the range used in most earlier studies, and the subtlety of many of the emotions used (as compared to the limited sets of basic emotions used in earlier work).

Inter-rater reliability of emotion judgments and believability and intensity ratings. For emotion judgments, we computed separate confusion matrices for all raters (including double answers when two answers were provided) and correlated the confusion profiles of each rater with each other rater. An average profile correlation per rater was computed as an agreement index. Mean and range for this index per rater group (after excluding two outliers) are shown in Table 3. These average profile correlations, ranging from .76 to .88 are extremely high, given the complex nature of the task, and demonstrate a large extent of agreement between raters in assigning emotion labels. It should be noted that this holds even in cases in which the actor intention was not accurately inferred, suggesting that the portrayals generally provide relative unequivocal messages even if the actor did not succeed in portraying a specific emotion but rather a close member of the family or a similar emotion (as shown by the lawful patterns of confusion shown in Table 5).

The reliability of the ratings on the quantitative scale intensity (four levels labeled 1 to 4 from the least intense to the most intense) and believability (continuous visual analog scale raging from 0 to 10) was estimated with average intraclass correlation coefficients (ICCs) for the raters who provided a complete set of ratings (1,260 ratings for the portrayals produced by all 10 actors). For believability, the average ICC varies between 0.63 and 0.69; for intensity, ICC varies between 0.84 and 0.90. Further detail can be found in the Supplementary Materials. Again, given the complexity and amplitude of the task, as well as the difficulty of defining believability as a dimension, these coefficients compare favorably to what can be expected in most ratings studies (see Rosenthal, 1987). The lower level of agreement for believability is accounted for by the high degree of subjectivity in defining and judging this quality.

Table 3 about here

Accuracy for Differences Between Core Emotions, Presentation Modalities, and Verbal Content Types

First, we will discuss the ratings for the 12 portrayals produced with baseline intensity (i.e., portrayals that are not regulated) and for core emotions produced by all actors. A four-way repeated measures ANOVA was computed on the accuracy data (defined as the proportion of raters who provided one correct answer). Within variables are: Modality (3 levels: audio, video, audio-video) × Emotion (12 levels: pride, joy, amusement, interest, pleasure, relief, hot anger, panic fear, despair, irritation, anxiety, sadness) × Verbal Content (3 levels: Sentence 1, Sentence 2, “aaa”) and Repetition (2 levels: Instance 1 and Instance 2). The descriptive results are shown in Table 4a.

Table 4 about here

The ANOVA showed main effects for Modality and Emotion. The accuracy is lowest (.42) for portrayals presented in audio-only, somewhat higher (.55) for video-only modality, and most accurate (.61) for audio-video modality. For differences among the 12 core emotions – independently of verbal content, presentation modality, and repetition – the average accuracy varies greatly, as shown in the last column of Table 4, between .36 for despair and .81 for panic fear. No main effect was found for the two other variables, indicating that neither repetition nor sentence type systematically affected accuracy. It is particularly remarkable that the overall accuracy for portrayals featuring solely a sustained vowel was as high as the accuracy for pseudo-speech portrayals, which were on average much longer and could potentially include more cues.

Figure 2 about here

Significant two-way interaction effects were found for Modality × Emotion, and for Emotion × Verbal Content. The former, illustrated in Figure 2, is due to the fact that some core emotions go against the general trend (audio < video < audio-video. Thus, for hot anger accuracy is slightly higher for video-only than for the other two modalities and for joy, pride, sadness, and anxiety, accuracy based on video only is at about the same level as in the audio-video modality. The Emotion × Verbal Content effect is due to some emotions (e.g., relief, panic fear, and amusement), the portrayals using “aaa” are better recognized than those for pseudo-speech sentences, independently of the expressive modality considered, the opposite being true for other emotions (e.g., sadness, pride).

The difference between pseudo-speech as used in Sentences 1 and 2 and the sustained aaa is theoretically interesting, as it may indicate the differential role of certain phonemic cues. In contrast, the two sentences were construed according to the same principles and only served to examine the effect of different vowel sequences. To test directly the effect of difference between the two pseudo-sentences, a repeated measures ANOVA with the same four factors but including only two levels (Sentence 1 and Sentence 2) for the variable verbal content was run. No significant effects involving sentence type were found, suggesting that the difference due to the different phonetic material used in the two sentences can be disregarded and that the effects are likely to be similar for phoneme sequences of similar construction. Obviously, we cannot rule out specific effects of using linguistically meaningful speech material. The absence of an effect for repetition also confirms the high stability of the inferences based on the actor portrayals.

Figure 3 about here

To summarize the results for the different emotions in a systematic fashion according to the underlying Valence x Arousal design of the corpus described in the introduction, a repeated measures ANOVA for the mean accuracy scores computed for the four quadrants of Table 1 was performed. The results did not show any difference for valence but a main effect for arousal, with high aroused emotions significantly better recognized (average accuracy .58) than low aroused emotions (.48). However, there was also a significant interaction between valence and presentation modality, indicating that the recognition of positive emotions might rely much more on visual cues than that of negative emotions (see Figure 3). When *only* audio cues are available, accuracy for positive emotions (.38) is lower than for negative emotions (.47). But when the portrayals are presented in the *audio-video* modality, accuracy for positive emotions (.64) increases more than that for negative emotions (.59), suggesting that the association of audio and visual cues is especially important in order to accurately recognize positive emotions. A three-way interaction between valence, display modality, and arousal (see Figure 3), shows that the difference between positive and negative emotions is imputable to the negative high aroused emotions (panic fear, hot anger, despair), which are better recognized than other emotions specifically when they are presented in the audio only modality, and to the negative low aroused emotions (anxiety, irritation, sadness), which are less well recognized when presented in audio-video modality. For the sake of economy, we do not describe four additional three-way interactions here because they are of minor interest (see Supplementary Materials for further details).

Accuracy for Differences Between Additional Emotions, Presentation Modalities, and Verbal Content Types

To reduce the total number of portrayals to a manageable size, six additional emotions were portrayed by only half of the actors: Actors 2, 4, 5, 9, and 10 portrayed admiration, disgust, and shame, while actors 1, 3, 6, 7, and 8 portrayed tenderness, contempt, and surprise. The mean accuracy scores are listed in Table 4b and plotted in Figure 4. Two separate repeated measures ANOVAs on those two subsets of data were performed (using the same four factors as before) to analyze the effects on differences in accuracy. In both subsets of data, the ANOVAs showed a significant main effect of modality, again with audio only less well recognized than those with a video component. A main effect of emotion for Group 2 can be attributed to the less accurate recognition of shame portrayals. Significant Emotion x Modality interactions suggest that for some emotions (especially tenderness, surprise, admiration, disgust), accuracy is relatively higher when the portrayals are presented with sound and picture (audio-video); with the accuracy decreasing when sound is absent (video-only; see Figure 4). An Emotion × Verbal Content interaction in Group 2 indicates that portrayals of disgust using the “aaa” are more accurately recognized than those for pseudo-speech sentences, an effect that is accentuated when the portrayals are presented in audio-only modality. As for the core emotions, there are no differences between the two types of sentences and for repetition, suggesting a high level of stability of the effects over successive instances when produced by the same actors in the same recording session. The detailed statistics for these effects can be found in the Supplementary Materials.

Figure 4 about here

Inter-emotion Confusions and Reports of Mixed Emotions for Core and Additional Emotions With Baseline Intensity

Confusion matrices, showing the proportion with which each category is selected for each portrayed emotion, were created separately for the three presentation modalities and for verbal content (Sentence 1 vs. vowel “aaa,” see Note for Table 4). The detailed tables for these six confusion matrices are available in Supplementary Materials. For the sake of economy, only the major confusions (defined as larger than 2 × chance level) are shown in Table 5, along with the proportion of correct answers. The proportion of correct answers is computed on the basis of the diagonals in the confusion matrices, i.e. including double answers as separate answers, with at least one incorrect answer when a double answer is provided. These proportions are by definition slightly lower than the accuracy figures used to compute the ANOVAs reported earlier (where an answer was considered correct if one of two alternatives was correct).

Table 5 about here

For some emotions with low recognition accuracy, the answers are spread over several categories, whereas for others the confusions are much more systematic. A particularly striking example is shame produced with a sustained "aaa" and presented in the audio-only modality. Only 3% of the answers went to the correct category, shame, the remainder being spread over many categories, including the category neutral (i.e., not emotional by our definition). This is not the case for all emotions with low recognition rates; for example, for admiration (in the audio "aaa" condition), the correct label admiration represents 19% of all answers, whereas the label pleasure represents 31% and the label relief 39% of all answers provided, and other labels are never or rarely used. For some emotions, symmetric confusion patterns are found (for both types of verbal content and for all presentation modalities); this, sadness is often judged as despair and vice versa. Other confusions are asymmetrical: hot anger is often categorized as irritation, and amusement as joy, but only a few confusions go into the other direction. The fact that the most frequent confusions are not necessarily reciprocal suggests that the categories are not simply equivalent or synonymous.

As shown in Table 5, there are systematic confusions that are modality specific (or at least more salient in some modalities), depending on whether audio or video information is available. This suggests that confusions may be partly based on lack of salient cues when only a single channel is available, and/or that it takes specific cues in a specific modality to recognize certain emotions.

While the confusion matrix provides very rich information and can be the source of important hypotheses for future research with respect to emotion similarities between and within families and the nature of the differentiating cues, a more detailed discussion would exceed the confines of this chapter which is mostly focused on reliability and validity of the corpus. The latter are confirmed by the fact that the confusions are generally meaningful and give rise to justifiable interpretations.

Accuracy for Portrayals Produced with Regulation Attempts (Modulations of Intensity and Masking)

Two separate repeated measures ANOVAs were computed with the accuracy data for the two subsets of emotions portrayed by different actors (see Method). The ANOVAs included four within factors: regulation (four levels: masked, less intense, baseline intensity, more intense); emotion (six levels: hot anger, despair, anxiety, amusement, interest, and pleasure in the first analysis; pride, joy, relief, panic fear, irritation, and sadness in the second analysis); modality (three levels: audio, video, audio-video); and repetition (two levels: Instance 1 and Instance 2). Repeated contrasts were computed to estimate the effect of the four regulations on recognition accuracy. Contrasts were defined on the basis of the hypothesis that the masked portrayals would be the least well recognized (because they are disguised) and that less intense emotion portrayals would be more subtle and therefore less accurately recognized than portrayals produced with baseline intensity or more intense emotion portrayals. Contrasts also tested the assumption that the more intense emotion portrayals would be more accurately recognized than would the portrayals with baseline emotional intensity, provided that more emotional intensity might result in more stereotypical portrayals.

Figure 5 about here

The analysis of those two subsets of data showed differences for emotion and display modality comparable to those described in the previous section. Regarding the influence of the regulations (masking the emotion, baseline intensity, and less and more intense emotions), there was a main effect of regulation in both subsets, shown in Figure 5. The contrasts showed that the masked portrayals were less accurately recognized than were other portrayals.

The differences between the three degrees of intensity go into the expected direction but do not reach significance. Statistical coefficients and further detail is provided in the Supplementary Materials.

Intensity Ratings for Portrayals Produced with Regulation Attempts (Modulations of Intensity and Masking)

An average intensity rating was computed for each portrayal. When a rater reported two emotion labels with different intensities for one portrayal, the highest intensity reported was retained. When a rater chose to indicate that a portrayal did not express an emotion or that it expressed an emotion not listed among the 15 alternatives proposed for each portrayal, he or she did not explicitly report an intensity level; such answers were therefore not used for the computation of the average intensity score.

We expected that the average intensity ratings would vary in accordance with the instructions provided to the actors regarding intensity regulations (baseline intensity, less intense and more intense emotion portrayals). To test this assumption, two separate repeated measures ANOVAs were computed on two subsets of data, as described in the preceding section. The masked portrayals were not included in this analysis because we did not expect those portrayals to be as accurately recognized as the other portrayals and made no assumptions regarding their emotional intensity. The ANOVAs included four within factors: Regulation (three levels: less intense, baseline intensity, more intense), Emotion (six levels: hot anger, despair, anxiety, amusement, interest, and pleasure in the first analysis; pride, joy, relief, panic fear, irritation, and sadness in the second analysis), Modality (three levels: audio, video, audio-video), and Repetition (two levels: Instance 1 and Instance 2). Repeated contrasts were computed to estimate the effect of the three intensity regulations on the average intensity rating. Contrasts were defined based on the hypothesis that the less intense emotion portrayals would be rated as less intense than portrayals produced with baseline intensity and that portrayals with more intensity would be rated as more intense than portrayals with baseline intensity.

Figure 6 about here

All main effects except the repetition were significant. Most importantly, the contrasts confirmed the expected differences for regulated portrayals in both groups (see Figure 6). The less intense emotion portrayals (2.53 in Group 1 and 2.46 in Group 2) were rated as less intense than were the portrayals produced with baseline intensity (2.73 and 2.71) and the more intense emotion portrayals (3.09 and 3.11) were indeed rated as more intense than were the portrayals produced with baseline intensity. Further details, statistical coefficients, and a plot of the means can be found in the Supplemental materials.

Believability Ratings for Portrayals Produced with Regulation Attempts (Modulations of Intensity and Masking)

The effects of different regulations (baseline intensity, less intense, more intense, masked), which was previously assessed for accuracy, were tested in the same way for the average ratings of believability computed for each portrayal in each display modality (audio, video, and audio-video). We repeated the statistical analyses described earlier (two independent repeated measures ANOVAs for two data subsets featuring different actors and different emotions), but the assumptions tested by the contrast analyses were different. We predicted that the masked portrayals would be rated as the least believable. This assumption relies on the instructions provided to the actors to disguise their emotional displays; this was thought to introduce conflicting cues that might result in awkward and less believable emotion portrayals. We furthermore speculated that the instruction to produce more intense emotional portrayals might result in overacting and consequently drive raters to perceive the more intense portrayals as less believable than the portrayals with baseline intensity. Finally, we also hypothesized that the instruction to produce less intense portrayals might have the opposite effect (i.e., prime the actors to produce more subtle and maybe also more realistic emotion portrayals).

The results are shown in Figure 7. The repeated measures ANOVA computed for the first group yielded significant contrasts showing that the masked portrayals were rated as less believable (average 6.3) than were the other portrayals (average 6.7 for more intense emotion portrayals and for baseline intensity; 6.6 for less intense emotion portrayals). There were no significant differences between degrees of intensity. The low level of believability of masked portrayals was confirmed for the second group, but here significant contrasts showed that the masked portrayals were not rated as significantly less believable (average 6.1) than were the more intense portrayals (average 6.3), whereas the portrayals with baseline intensity (average 6.7) were rated as more believable than were both the less intense emotion portrayals (average 6.4), and the more intense emotion portrayals (average 6.3).

Figure 7 about here

Relationships Between Accuracy, Believability, and Intensity Ratings

We computed correlations between accuracy, average believability, and average intensity ratings for all portrayals selected in the corpus (*N* = 1260). Separate correlations were computed for the three presentation modalities (see Table 6). The relationship between accuracy (proportion of raters who recognize the emotion portrayed in a given audio or video recording) and believability (average ratings for the authenticity of the portrayals) is interesting in several respects. There are explicit speculations about acted emotion portrayals, to the effect that acted emotion portrayals that are highly recognizable are also very stereotypical and would not create an authentic or realistic impression. The significant correlations between accuracy and believability suggest the opposite interpretation. The more readily recognizable a portrayal is, the more believable (authentic or realistic) it was rated in our study. Intensity ratings appear to be correlated with both accuracy and believability, indicating that more extreme portrayals (that are rated as expressing strong emotions) are not only better recognized than less extreme portrayals, but they are also perceived to be more authentic or more realistic.

Table 6 about here

Discussion and outlook

Validation of the corpus

These results suggest that the GEMEP corpus has been successfully validated, given the satisfactory degree of inter-rater agreement (reliability) and the high level of accuracy. It is important to note that a complete set of portrayals in all conditions of the corpus design were rated – a total of 1260 portrayals. The degree of accuracy found compares very favorably with established tests of emotion recognition. Table 7 shows a comparison of the GEMEP results with those of five tests obtained in a recent study (Bänziger et al., 2009): the MERT (The Multimodal Emotion Recognition Test; Bänziger et al.), the PONS (Profile of Nonverbal Sensitivity; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979), the DANVA (Diagnostic Analysis of Nonverbal Accuracy, Nowicki & Duke, 1994), the ERI (Emotion Recognition Index; Scherer & Scherer, 2008), and the JACFEE (Japanese and Caucasian Facial Expressions of Emotion; Biehl et al., 1997). One source of incompatibility between tests stems from differing numbers of response alternatives on the answer sheet. To render the accuracy percentages comparable across tests despite differential answer formats, we computed the one-sample effect size estimator called the Proportion Index, or pi (Hall, Andrzejewski, Murphy, Schmid Mast, & Feinstein, 2008; Rosenthal & Rubin, 1989), shown in Table 7. Pi converts any mean accuracy that originates as a proportion, no matter how many response options each item had, to its equivalent proportion were it to have been based on two options.

Table 7 about here

The pi values for the GEMEP corpus are at least as high and in some cases higher than the pi values for the emotion recognition tests. This is remarkable if one considers that the items used in the tests have generally been carefully selected for accuracy from a larger set of portrayals, whereas the GEMEP values are based on a complete set of portrayals in the corpus that have not been selected earlier for accuracy. It is important to note that this holds not only for the so-called basic emotions, but also for the large number of more subtle, and much less studied, emotions that are included in the corpus.

Another aspect of validation is to establish the believability of the portrayals, which essential to use the corpus as a valid source for stimulus presentation in studies of the perception of and inference from emotional expression. Actors were instructed to produce expressions that make an authentic impression on the receivers of their performance – as they would attempt to do on the stage. The judges in this study rated the believability of the emotion portrayals defined as the success of the actor to have produced an authentic and plausible emotional impression. While the overall level of agreement on this quality is lower than for intensity and category judgments (as is to be expected on the basis of individual difference in the evaluation of this highly subjective construct), there is still is substantial agreement suggesting that this dimension can be rated with sufficient reliability. As a consequence, the believability ratings for individual ratings can be used (just as the differential accuracy scores) to select specific items out of the total set for specific subsets of stimulus material.

Another aspect of validity is the assurance that the impressions produced by the portrayals are stable in the sense of not depending too strongly on the nature of the vocal utterances used for the portrayals or on random effects. The results reported above show that the ratings are stable over repetitions and over two different nonlinguistic sentences, suggesting that portrayals can be chosen without having to consider these factors. However, the factor *utterance type*, that is, nonlinguistic sentences versus affect bursts, does make a difference and thus researchers need to determine which utterance type fits best for their respective research purpose. Most likely, the affect burst provides a more primitive and less social instance of EE because it is not influenced by phonological, syntactic, and prosodic factors and may well occur when the sender is alone. In contrast, the sentence-like utterances can be expected to be determined in part by these linguistic factors and to be closer to a typical utterance in social interaction.

An innovative feature of the GEMEP corpus is the attempt to study masking and variations in the intensity of expression, especially given the frequent critique that actor portrayals are too “stereotypical”. Importantly, even though recognition accuracy is lower for the masked stimuli as compared to the non-masked ones, they are still recognized at an accuracy level that is higher than chance (Figure 5), although believability is significantly lower than that for normal portrayals at different intensities (Figure 7). Apart from this case, the results show that the regulated intensity levels of the portrayals were indeed correctly perceived by the raters, as shown by the intensity ratings in Figure 6. This means that the corpus can be used to select portrayals at different intensities to systematically study the effect on distal features and emotion inference and attribution.

Modality by Emotion Effects on Accuracy

The empirical results for recognition differences between emotions across modalities provide a first basis for the development of hypotheses about the type of distal and proximal cues that may be involved in the communication of emotion. As one would have expected, the audio-visual condition produces the greatest degree of accuracy, given that it provides all available cues. In comparing single channels, as expected on the basis of an earlier review of channel studies (Scherer, 1999), the emotions in the audio modality are less well recognized than they are in the video modality. However, the data show (Figures 2 and 4) that this result is mostly due to a few emotions in which audio accuracy is low and video accuracy appreciably higher. These cases seem to be limited to a few emotion families such as disgust and contempt, which confirms earlier findings. One explanation is that disgust is usually a very brief emotion during which nothing is spoken (Banse & Scherer, 1996). Thus, it may not be surprising that actors find it difficult to convey the emotion in a sentence-like utterance. An interesting finding is that the accuracy proportion jumps from .12 for the sentence case to .59 in the affect burst case (see Table 4), indicating the existence of specific vocal affect emblems (see Scherer, 1994a). Similarly, contempt may be an attitude toward another person that is shown in the face but that rarely colors interactive speech for a longer period. The other set of emotions in which the audio modality is clearly disadvantaged are the positive emotions of interest, joy, and pride. In the case of interest, little voice change seems to occur from neutral (this being one of the few cases in which the tendency to use the neutral label is strong; see Table 5). In the case of joy and pride, a number of unambiguous facial signs – including the smile – are apparent, whereas specific cues do not seem to occur in the voice. We find it interesting that pride is systematically confused with irritation in the voice. In the case of sadness and despair, accuracy is low in all modalities, mainly because of the systematic symmetric confusions between the two emotions.

Future development of the GEMEP corpus

As outlined in CHAPTER 3.2, one of the essential purposes of actor portrayal studies is to determine the distal and proximal cues and cue utilization in the process of emotion communication. Thus, one important direction of further research is the extraction, coding, or annotation of the behavioral features that distinguish the expression of emotions. The techniques to annotate and analyze these features are extremely costly and time-consuming, and it would thus be unrealistic to analyze all 1,260 stimuli. Similarly, for future studies it would be difficult to have to deal with such a large number of portrayals. In consequence, the selection of a core set is required.

Core set. We decided to identify, on the basis of the ratings reported here, a subset of portrayals, representative of all the emotions and actors, that have received high believability ratings and have a satisfactory level of accuracy, showing that most observers will unambiguously classify them according to portrayal intention. This core set was subjected to another extensive rating study with a much larger number of raters, as well as a categorical response scheme and dimensional ratings (in two separate subgroups of ratings). The results of this study are currently being prepared for publication.

Vocal analysis. In the vocal domain, the state of the art is the extraction of acoustic parameters by using digital signal analysis procedures (see Banse & Scherer, 1996; Juslin & Scherer, 2005; Scherer, Johnstone, & Klasmeyer, 2003). Vocal parameter extraction and analysis has been performed for the core set, and an article reporting the results has been submitted for publication (Goudbeek & Scherer, 2009). Current work (with J. Sundberg) is focused on a microanalysis of the vocal affect bursts.

Facial analysis. The state-of-the-art instrument to objectively determine the facial movements in expression is the Facial Action Coding System (FACS, Ekman & Friesen, 1978). The GEMEP core set is currently FACS coded by certified coders, which constitutes a very time-consuming activity. First results on disambiguating subtly different positive emotions are reported in an article by Mortillaro, Mehu, and Scherer (2009). Similar work is under way to study the patterns of action units (AUs) that differentiate families of negative emotions. Because the core set sequences are dynamically coded for onset, apex, and offset of each of approximately 40 AUs, fine-grained analyses of the sequential emergence of AUs and other dynamical aspects of facial expression are currently being performed (Krumhuber, Mehu, & Scherer, 2009). The results will provide a test of Scherer's assumption of sequential unfolding of facial expression as driven by appraisal checks (Aue & Scherer, 2008; Delplanque et al., 2008; Scherer, 1992, 2001, 2009).

Gesture and posture annotation. The study of EE via gesture and posture has been remarkably neglected in the field (but see Wallbott, 1998). A new comprehensive gesture and posture coding system has been developed recently in Geneva and the GEMEP core set will be coded by using the same time line as for face and voice. Particular emphasis will also be placed on head movements.

Multimodal synchronization. Because all modalities are coded on the same time line, it will be possible to examine the coherence or synchronization between these systems for the different emotions. As suggested by Scherer (1984, 2005, 2009), a high degree of subsystem synchronization can be seen as the hallmark for the presence of an emotion. Special attention will be paid to the role of synchronization in perceived emotional authenticity.

Regressing behavior on observer ratings. As mentioned at the outset, one of the main purposes of this research program is to empirically investigate the Brunswikian lens model in the context of emotion communication. In consequence, much of our work is based on correlating the behavioral data with the subjective ratings to determine the cues used by the observers in their inference and attribution. The crowning piece of this type of analysis is a path analysis or structural modeling to map the data into a Brunswikian lens model.

Using the corpus as stimulus material. An important asset of acted emotion portrayals lies in the absence of contextual cues or variability attached to an emotion-eliciting situation. Unlike “natural” (spontaneously occurring) EEs, actor portrayals with standard verbal content contain only nonverbal cues to emotions. This allows us to use them to test a variety of hypotheses. Hence, much of the ongoing work uses the GEMEP portrayals as systematic and standardized stimulus material in psychological and neuroscientific studies. Ethofer, Van De Ville, Scherer, and Vuilleumier (2009) recently used audio GEMEP portrayals to study the decoding of emotional information in voice-sensitive cortices. The portrayals are currently used in several neuroscience applications. We plan the development of several adaptive tests of emotion recognition for research use, evaluation of emotional competence, and diagnosis of neurological damage. Another area in which the GEMEP corpus is of great utility is in the area of affective computing, for example, the development of dynamic, sequential facial synthesis (Roesch et al., 2009). Similarly, the GEMEP portrayals might be used to test the effect of contextual information on the interpretation of the portrayed emotions by providing various explanations along the portrayals (e.g., by allocating various meanings to the pseudo-speech sentences pronounced by the actors).

In sum, the GEMEP corpus is a comprehensive, sophisticated, and valid new instrument for research on emotional expression in many different areas such as psychological research on perception and communication of emotion, neuroscience research on the brain structures and circuits underlying emotion expression processing, or work in affective computing (see CHAPTER 6.2 and CHAPTER 3.2). The GEMEP corpus is shared with these research communities (see Footnote 1 for details) and the complete data base with all pertinent annotations will be made available once parameter extraction is finished.

Acknowledgement

The research reported here has been supported by the German and the Swiss National Research Foundations, the Swiss Center for Affective Sciences, and the FP6 IST European Network of Excellence Humaine. The authors gratefully acknowledge important contributions by Steve Binggeli in the construction of the corpus and collection of ratings and by Nele Dael, Martijn Goudbeek, Marc Mehu, and Marcello Mortillaro, in the final data gathering stages and in valuable comments all along.

References

Aue, T., & Scherer, K. R. (2008). Appraisal-driven somatovisceral response patterning: Effects of intrinsic pleasantness and goal conduciveness. *Biological Psychology, 79,* 158–164.

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

Bänziger, T., Grandjean, D., & Scherer, K.R. (2009). Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion, 9,* 691-704.

Bänziger, T., & Scherer, K. R. (2007). Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In A. Paiva, R. Prada, & R. Picard (Eds.), *Affective computing and intelligent interaction 2007: Lecture notes in computer science: Vol. 4738* (pp. 476-487). Berlin, Germany: Springer-Verlag.

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., et al. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21*, 3-21.

Delplanque, S., Grandjean, D., Chrea, C., Aymard, L., Cayeux, I., Margot, C., et al. (2009). Sequential unfolding of novelty and pleasantness appraisals of odors: Evidence from facial electromyography and autonomic reactions. *Emotion, 9*, 316-328.

Ekman, P. (1999). Basic Emotions. In T. Dalgleish and T. Power (Eds.) *The Handbook of Cognition and Emotion* (pp. 45-60). Sussex, U.K.: John Wiley & Sons, Ltd.

Ekman, P., & Friesen, W. (1976). Pictures of facial affect. Palo Alto: Consulting Psychologists Press.

Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of emotional information in voice-sensitive cortices. *Current Biology*, 19(12), 1028-1033.

Goeleven, E., De Raedt, R., Leyman, L., and Verschuere, B. (2008). The Karolinska directed emotional faces: a validation studies. *Cognition and Emotion, 22(6):*1094–1118.

Gosselin, P., Kirouac, G., & Doré, F. Y. (1995). Components and recognition of facial expression in the communication of emotion by actors. *Journal of Personality and Social Psychology, 68,* 1-14.

Goudbeek, M., & Scherer, K. R. (2009). Beyond arousal: Valence and potency/control in the vocal expression of emotion. Manuscript submitted for publication.

Hall, J. A., Andrzejewski, S. A., Murphy, N. A., Schmid Mast, M., & Feinstein, B. A. (2008). Accuracy of judging others' traits and states: Comparing mean levels across tests. *Journal of Research in Personality, 42,* 1476–1489.

Hawk, S. T., Van Kleef, G. A., Fischer, A. H., & Van der Schalk, J. (2009). "Worth a thousand words": Absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion, 9,* 293-305.

Izard, C. E. (1991). *The psychology of emotions*. New York: Plenum Press.

Johnson, W.F., Emde, R.N., Scherer, K. R., Klinnert, M.D. (1986). Recognition of emotion from vocal cues. Archives of General Psychiatry, 43, 280-283.

Juslin, P.N., & Scherer, K. R. (2005). Vocal expression of affect. In J. A. Harrigan, R. Rosenthal, & K. Scherer, (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 65-135). Oxford, UK: Oxford University Press.

Krumhuber, E., Mehu, M., & Scherer, K. R. (2009). *Dynamic unfolding of emotions portrayed by actors*. Manuscript in preparation.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from the Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, Stockholm, Sweden.

Maurage, P., Joassin, F., Philippot, P., & Campanella, S. (2007). A validated battery of vocal emotional expressions. *Neuropsychological Trends, 2*, 63-74.

Mortillaro, M., Mehu, M., & Scherer, K. R. (2009). *Subtly different positive emotions can be distinguished by dynamic, appraisal-driven facial expressions.* University of Geneva. Manuscript submitted for publication.

Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy. *Journal of Nonverbal Behavior, 18*, 9-35.

Roesch, E.B, Tamarit, L., Reveret, L., Grandjean, D., Sander, D., Scherer, K. R. (in press). FACSGen: A Tool to Synthesize Emotional Facial Expressions through Systematic Manipulation of Facial Action Units. Journal of Nonverbal Behaviour.

Rosenthal, R. (1987). Judgment studies: Design, analysis, and meta-analysis. Cambridge: Cambridge University Press.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test.* Baltimore: John Hopkins University Press.

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin, 106*, 332–337.

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology, 8*, 467-487.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 293-318). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1992). What does facial expression express? In K. Strongman (Ed.), *International review of studies on emotion* (Vol. 2, pp. 139-165). Chichester, England: Wiley.

Scherer, K. R. (1994). Affect bursts. In S. van Goozen, N.E. van de Poll, & J.A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 161-196). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1999). Universality of emotional expression. In D. Levinson, J. Ponzetti, & P. Jorgenson (Eds.), *Encyclopedia of human emotions* (Vol. 2, pp. 669-674). New York: Macmillan.

Scherer, K. R. (2001). Appraisal considered as a process of multi-level sequential checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 92-120). New York: Oxford University Press.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*, 693-727.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion, 23*(7), 1307-1351.

Scherer, K. R., & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion, 7*, 113-130.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer , H. Goldsmith (Eds.), *Handbook of the affective sciences* (pp. 433-456). New York: Oxford University Press.

Scherer, K. R., & Scherer, U. (2008). *Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the Emotion Recognition Index (ERI).* Manuscript submitted for publication.

Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, *28*, 879–896.

Author Note

Tanja Bänziger and Klaus R. Scherer, Swiss Centre for Affective Sciences, University of Geneva, Switzerland.

The research reported here has been supported by the German and the Swiss National Research Foundations, the Swiss Center for Affective Sciences, and the FP6 IST European Network of Excellence Humaine. The authors gratefully acknowledge important contributions by Steve Binggeli in the construction of the corpus and collection of ratings and by Nele Dael, Martijn Goudbeek, Marc Mehu, and Marcello Mortillaro, in the final data gathering stages and in valuable comments all along.

Footnotes

[1] Supplementary materials: The complexity of the GEMEP corpus, along with the richness of the rating studies, prevents us from including all the available data in one chapter. Nevertheless, we decided to provide the research community with this descriptive data. The data is published online in *The GEMEP Primer* and can be downloaded from www.affective-sciences.org/gemep. The primer is intended to be the reference guide that includes detailed information on the corpus and its creation as well as all the available data generated for the GEMEP corpus. In consequence, it will be continually enriched with newer data

Table 1

Selection of Emotions Portrayed

	Valence	
Arousal	Positive	Negative
High	Elation (joy)	Hot anger (rage)
	Amusement	Panic fear
	Pride	Despair
Low	Pleasure	Cold anger (irritation)
	Relief	Anxiety (worry)
	Interest	Sadness (depression)

Note. Additional states: shame, surprise, admiration, disgust, contempt, tenderness.

Table 2

Definitions of Emotions Portrayed

Emotion	Definition
Admiration	Amazement at the extraordinary qualities of a person, a landscape, or a work of art
Amusement	Roaring with laughter at something that is very funny
Anger	Extreme displeasure caused by someone's stupid or hostile action
Tenderness	Being moved by a touching action, behavior, or utterance
Disgust	Revulsion when faced with an unpleasant object or environment
Despair	Distress at a life problem with no solution, together with an unwillingness to accept the situation
Pride	Feeling of triumph following a success or a personal achievement (one's own or that of someone close)
Shame	Self-esteem shaken by an error or clumsiness for which one feels responsible
Anxiety (worry)	Fear of the consequences of a situation that could be unfavorable for oneself or someone close
Interest	Being attracted, fascinated, or having one's attention captured by a person or a thing
Irritation	Experiencing displeasure at something or someone while still remaining calm
Joy (elation)	Feeling transported by a fabulous thing that occurred unexpectedly
Contempt	Disapproval of the socially or morally reprehensible conduct of another person
Panic fear	Being faced with an imminent danger that threatens our survival or physical well-being
Pleasure (sensual)	Experiencing an extraordinary feeling of well-being and sensual delight
Relief	Feeling reassured at the end or resolution of an uncomfortable, unpleasant, or even dangerous situation
Surprise	Being abruptly faced with an unexpected and unusual event (without positive or negative connotation)
Sadness	Feeling discouraged by the irrevocable loss of a person, place, or thing

Table 3

Accuracy for the Three Presentation Modalities and Average Inter-rater Profile Correlations

	Audio-Video	Audio	Video
Number of raters	23	23	25
Average accuracy (%)	57	38	50
Range (%)	35	29	22
Average profile correlation r	.89	.76	.87
Range of r	.14	.23	.09

Table 4

Accuracy score means for All Emotions By Two Verbal Contents and Three Display Modalities

Accuracy			Audio-Video		Audio		Video		Grand total
Valence	Arousal	Target emotion	Sent 1	aaa	Sent1	aaa	Sent 1	aaa	
a) Core emotions									
Positive	High	pri	0.64	0.50	0.24	0.10	0.57	0.35	0.40
		joy	0.70	0.55	0.35	0.20	0.67	0.64	0.52
		amu	0.72	0.87	0.57	0.78	0.69	0.74	0.73
	High total		0.69	0.64	0.39	0.36	0.64	0.58	0.55
	Low	int	0.52	0.56	0.26	0.30	0.44	0.52	0.43
		ple	0.61	0.56	0.31	0.40	0.39	0.38	0.44
		rel	0.77	0.90	0.49	0.73	0.64	0.76	0.71
	Low total		0.63	0.67	0.35	0.48	0.49	0.55	0.53
Positive total			0.66	0.65	0.37	0.42	0.57	0.57	0.54
Negative	High	hot	0.69	0.76	0.67	0.72	0.76	0.80	0.73
		pan	0.79	0.97	0.66	0.81	0.66	0.97	0.81
		des	0.43	0.48	0.31	0.33	0.25	0.35	0.36
	High total		0.64	0.74	0.55	0.62	0.55	0.71	0.63
	Low	irr	0.64	0.59	0.51	0.31	0.50	0.48	0.50
		anx	0.57	0.54	0.40	0.29	0.52	0.45	0.46
		sad	0.43	0.26	0.45	0.23	0.43	0.41	0.37
	Low total		0.54	0.46	0.45	0.28	0.49	0.45	0.44
Negative total			0.59	0.60	0.50	0.45	0.52	0.58	0.54
Total			0.63	0.63	0.44	0.43	0.54	0.57	0.54

b) Additional emotions								
	adm	0.54	0.61	0.39	0.23	0.39	0.51	0.44
	sha	0.22	0.24	0.11	0.03	0.27	0.29	0.19
	con	0.66	0.54	0.25	0.10	0.61	0.57	0.45
	ten	0.45	0.70	0.30	0.22	0.27	0.53	0.41
	dis	0.76	0.98	0.12	0.59	0.50	0.77	0.62
	sur	0.56	0.73	0.33	0.47	0.33	0.59	0.50
Total		0.53	0.63	0.25	0.27	0.40	0.54	0.44
Grand total		0.61	0.63	0.40	0.40	0.51	0.57	0.52

Note. For space efficiency, the table displays only the values for Sentence 1 as no significant difference was found between Sentence 1 and Sentence 2.

Abbreviations: Sent1 = Sentence 1; aaa = sustained vowel “aaa”; pri = pride; joy = joy; amu = amusement; int = interest; ple = pleasure; rel = relief; hot = hot anger; pan = panic fear; des = despair; irr = irritation; anx = anxiety; sad = sadness; adm = admiration; sha = shame; con = contempt; ten = tenderness; dis = disgust; sur = surprise. In this table, accuracy means the average proportion of raters choosing the correct category.

Overall ANOVA main effects for modality, $F(2, 16) = 93.19$, $p < .001$, $\eta^2 = .92$, and emotion, $F(11, 88) = 19.84$, $p < .001$, $\eta^2 = .71$. All differences between modality levels are significant in post hoc tests with Bonferroni adjustment. two-way interaction effects for Modality × Emotion, $F(22, 176) = 6.66$, $p < .001$, $\eta^2 = .45$, and for Emotion × Verbal Content, $F(22, 176) = 5.76$, $p < .001$, $\eta^2 = .42$. No three-way interaction effects beyond what could be expected by chance.

Valence x Arousal ANOVA: main effect for arousal, $F(1, 9) = 53.90$, $p < .001$, $\eta^2 = .86$. ; two-way interaction between valence and display modality, $F(2, 18) = 19.96$, $p < .001$, $\eta^2 = .69$; three way effects, Modality × Valence × Verbal Content, $F(4, 36) = 3.43$, $p = .018$, $\eta^2 = .28$; Verbal Content × Repetition × Arousal, $F(2, 18) = 4.53$, $p = .025$, $\eta^2 = .34$; Verbal Content × Valence × Arousal, $F(2, 18) = 37.45$, $p < .001$, $\eta^2 = .81$; and Repetition × Valence × Arousal. $F(2, 18) = 17.54$, $p < .001$, $\eta^2 = .66$.

Table 5

Proportion of Answers Going to the Target Category and Major Confusions (>.125)

	AV-sent1		AV-aaa		A-sent1		A-aaa		V-sent1		V-aaa	
	% target	maj. conf.	% target	maj. conf.	% target	maj. conf.	% target	maj. conf.	% target	maj. conf.	% target	maj. conf.
pri	.56		.41	joy .18	.20	irr .24	.09	irr .16	.49	joy .18	.31	joy .34
joy	.60		.47	ple .15	.26		.16	amu .15 & pan .14	.56		.55	
amu	.68	joy .17	.81		.44	joy .18	.63	joy .26	.62	joy .19	.67	joy .19
int	.46		.47		.22	neu .13	.24	ple .15 & rel .18	.39	irr .18	.48	
ple	.53	rel .17	.49	rel .29	.26	rel .17	.32	rel .36	.33	rel .20	.33	rel .29
rel	.69		.82		.42		.60	ple .22	.56	ple .14	.70	ple .15
hot	.64	irr .31	.71	irr .25	.51	irr .35	.59	irr .27	.67	irr .29	.75	irr .19
pan	.70	anx .18	.94		.52	anx .20	.71		.56	anx .20	.89	
des	.34	anx .17 & sad .20	.41	pan .21 & sad .21	.23	anx .19 & sad .18	.26	pan .26 & sad .17	.20	anx .20 & sad .23	.29	pan .15 & sad .37
irr	.58	hot .13	.54	hot .13	.43	hot .13	.26		.45	hot .16	.43	hot .15
anx	.51	pan .13	.47	pan .24	.34	irr .13	.25	pan .15	.47		.40	pan .15

sad	.38	irr .42	.24	des .26	.38	des .20	.19	rel .17 & des .17	.37	des .36	.37	des .32
adm	.48	ple .14	.52	ple .14 & rel .15	.33		.19	ple .31 & rel .39	.35	int .18	.45	rel .16
sha	.19	des .29	.21	des .15 & anx .18	.10	sad .16	.03	rel .15 & neu .14	.24	des .20	.26	des .15 & anx .15
con	.56	irr .14	.48	irr .19	.21		.08	ple .20 & rel .23 & irr .15	.54	irr .13	.51	
ten	.40	ple .24	.65	ple .15	.24	ple .15	.19	ple .14 & rel .13	.23	joy .15 & amu .14 & ple .24	.49	ple .13
dis	.70		.98		.10	sad .23	.55		.43	sad .25	.71	sad .20
sur	.46	anx .25	.60		.26	anx 15	.40	pan .15	.27	anx .29	.53	

Note. AV = audio-video; A = audio; V= video; sent1 = Sentence 1; aaa = sustained vowel “aaa”; maj. conf. = major confusions; pri = pride; joy = joy; amu = amusement; int = interest; ple = pleasure; rel = relief; hot = hot anger; pan = panic fear; des = despair; irr = irritation; anx = anxiety; sad = sadness; adm = admiration; sha = shame; con = contempt; ten = tenderness; dis = disgust; sur = surprise; neu = neutral.

Table 6

Inter-correlations Between Accuracy Proportions, Intensity, and Believability Ratings for Three Modalities

	Accuracy	Believability
Audio		
Believability	.483	
Intensity	.570	.518
Video		
Believability	.299	
Intensity	.573	.185
Audio-video		
Believability	.382	
Intensity	.597	.472

Note. All correlations are significant at $p < .001$ (two-tailed), $N = 1,260$.

Table 7

Comparison of the Proportion Index (pi) for Accuracy Scores in Different Tests of Emotion Recognition (for Three Modalities)

Test	Audio-Video	Audio (voice)	Video (face)
GEMEP	0.96	0.90	0.95
MERT	0.95	0.90	0.95
PONS	0.84	0.62	0.81
DANVA		0.88	0.94
ERI		0.88	0.92
JACFEE			0.95

Note. GEMEP = Geneva Multimodal Emotion Portrayal; MERT = Multimodal Emotion Recognition Test; PONS = Profile of Nonverbal Sensitivity; DANVA = Diagnostic Analysis of Nonverbal Accuracy; ERI = Emotion Recognition Index; JACFEE = Japanese and Caucasian Facial Expressions of Emotion.

Figure Captions

Figure 1. Still frames illustrating the three camera angles used in the video recording of the actor portrayals.

Figure 2. Accuracy for modalities and core emotions with baseline intensity. pri = pride; joy = joy; amu = amusement; int = interest; ple = pleasure; rel = relief; hot = hot anger; pan = panic fear; des = despair; irr = irritation; anx = anxiety; sad = sadness.

Figure 3. Accuracy for modalities, valence, and arousal with baseline intensity.

Figure 4. Accuracy for additional emotions and different modalities. ten = tenderness; con = contempt; sur = surprise; adm = admiration; dis = disgust; sha = shame.

Figure 5. Accuracy for types of regulation in two groups of actors portraying different emotions.

Figure 6. Intensity ratings for different emotions and types of regulation in two groups of actors portraying different emotions. amu = amusement; hot = hot anger; des = despair; anx = anxiety; int = interest; ple = pleasure; pri = pride; irr = irritation; joy = joy; pan = panic fear; rel = relief; sad = sadness.

Figure 7. Believability ratings for different emotions and types of regulation in two groups of actors portraying different emotions.

Figure 1

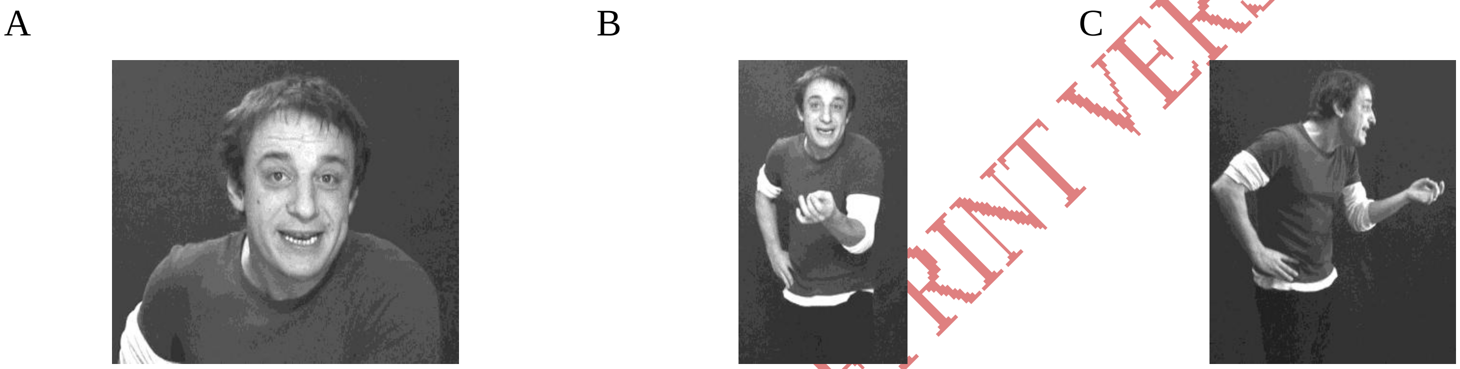


Figure 2

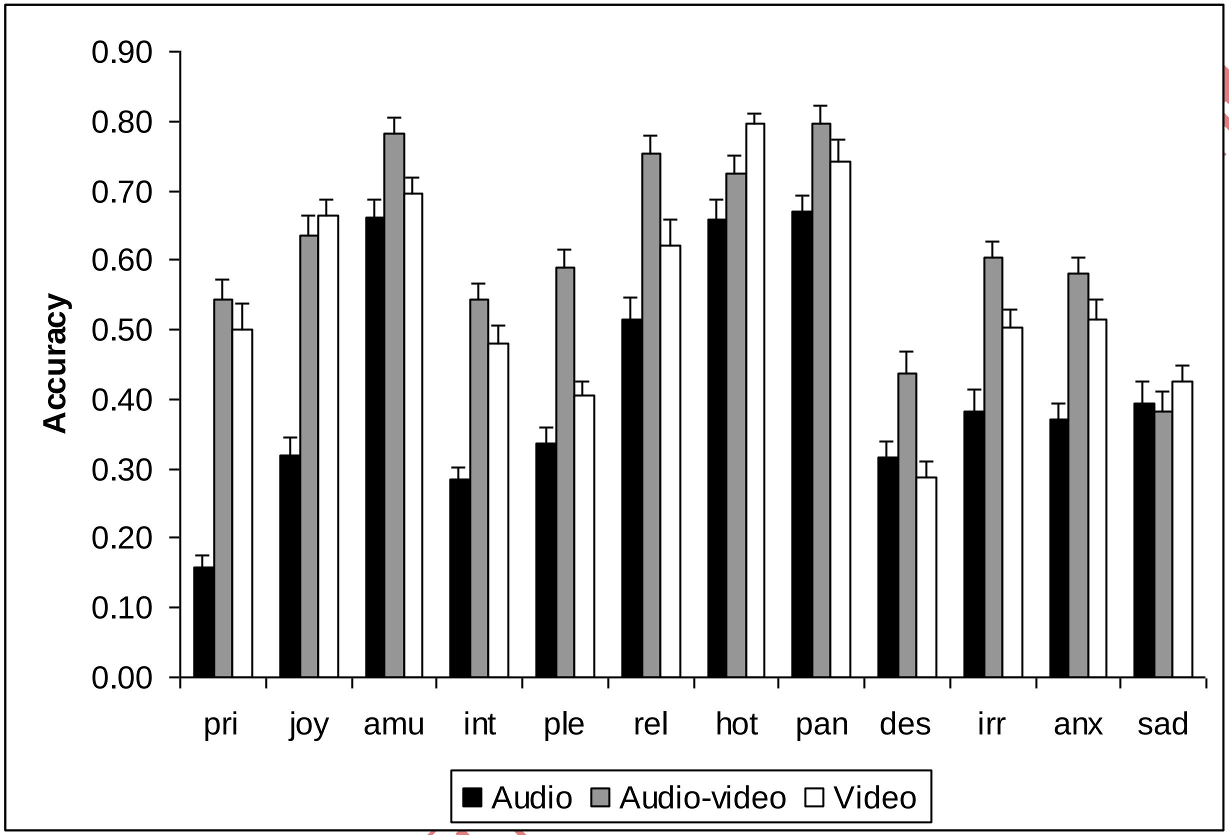

0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Accuracy
pri
joy
amu
int
ple
rel
hot
pan
des
irr
anx
sad
■ Audio
■ Audio-video
□ Video

Figure 3

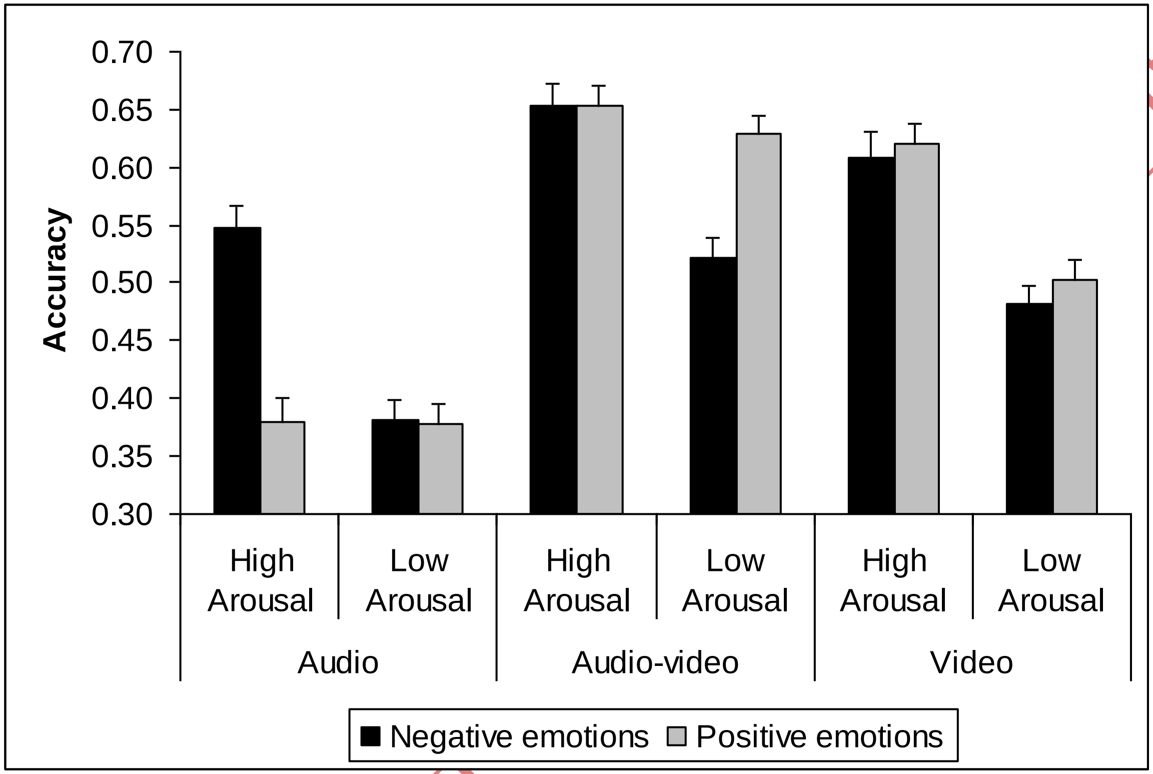

0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
Accuracy
High Arousal
Low Arousal
High Arousal
Low Arousal
High Arousal
Low Arousal
Audio
Audio-video
Video
■ Negative emotions
□ Positive emotions

Figure 4

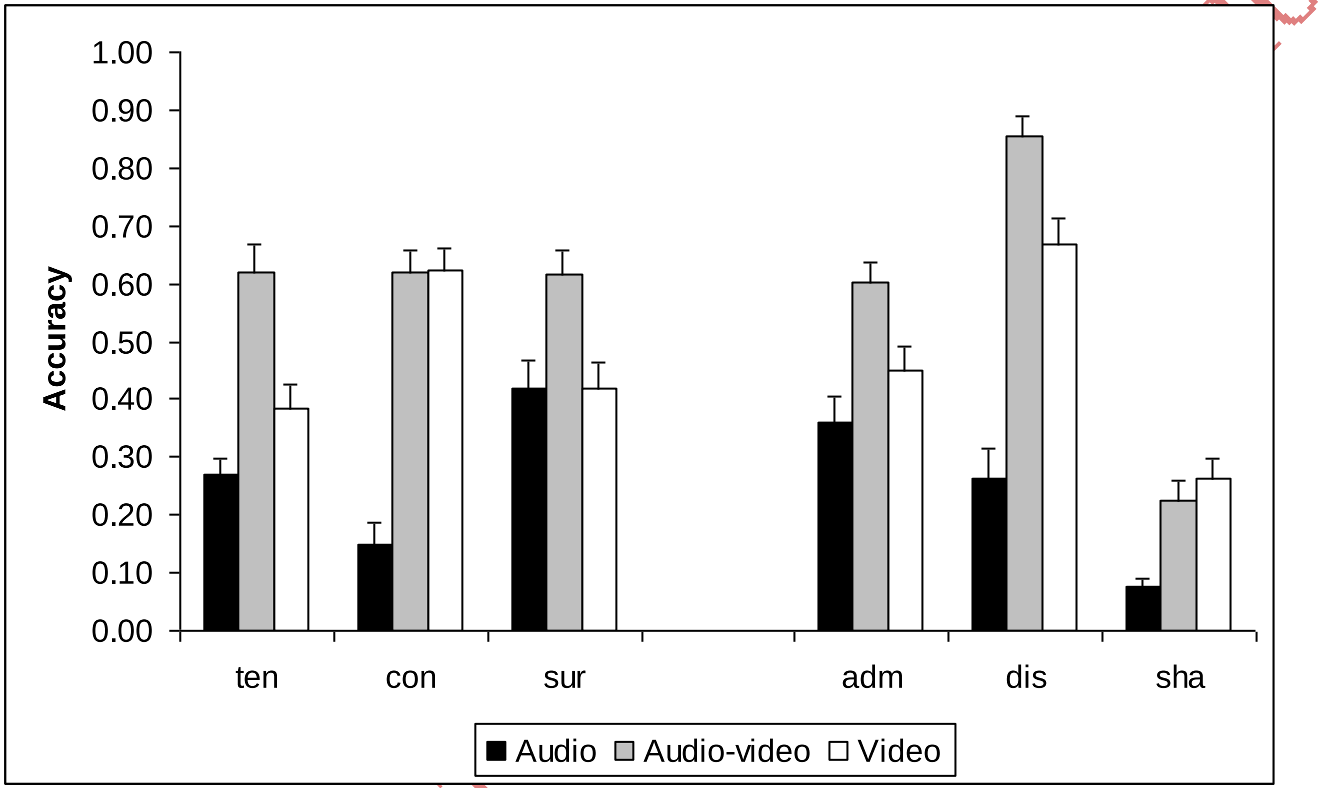

Accuracy
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
ten
con
sur
adm
dis
sha
■ Audio ▨ Audio-video □ Video

Figure 5

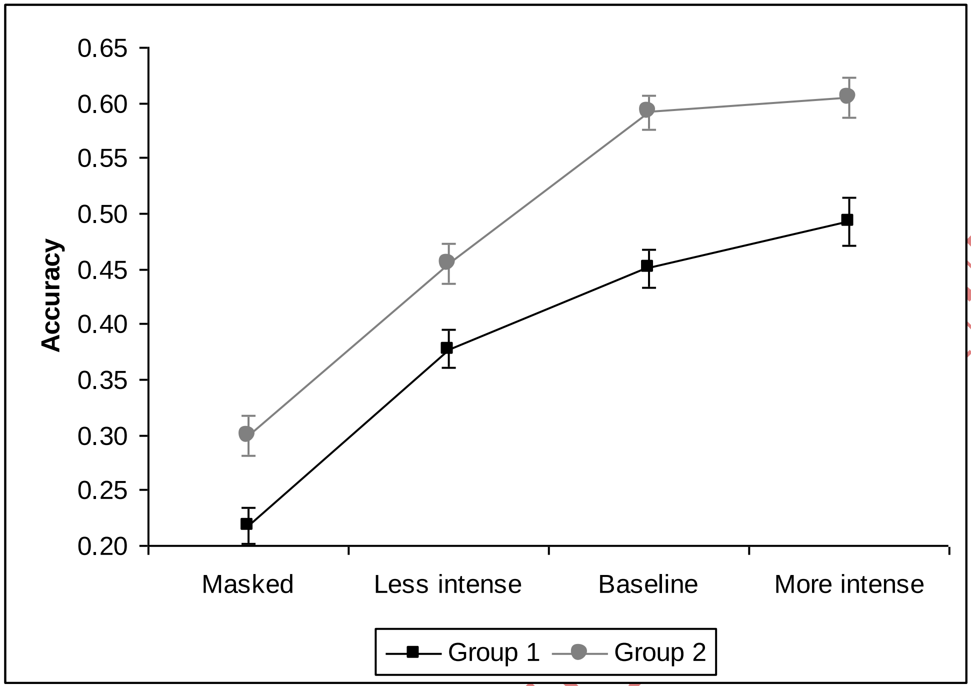

Accuracy
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
Masked
Less intense
Baseline
More intense
Group 1
Group 2

Figure 6

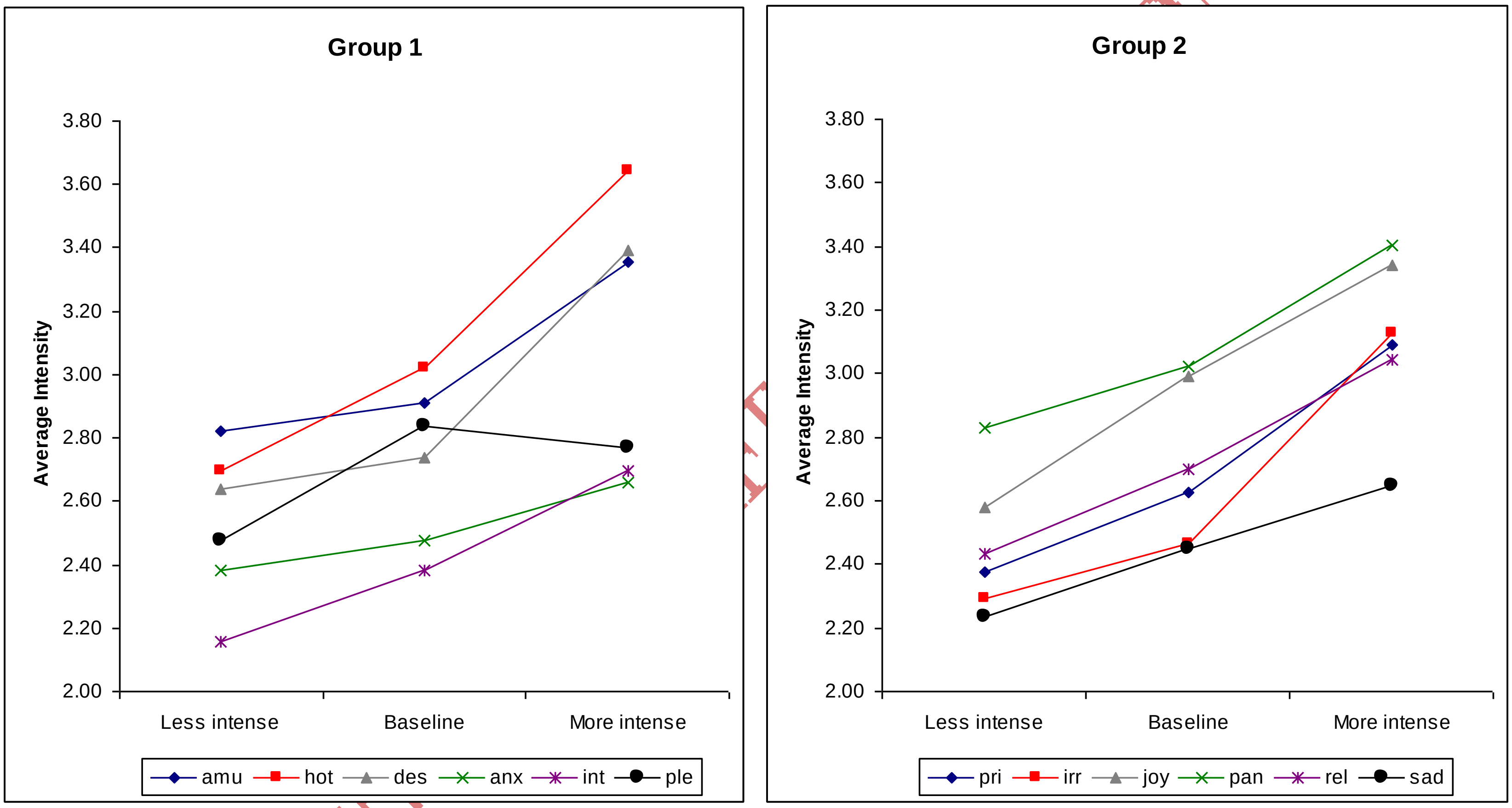
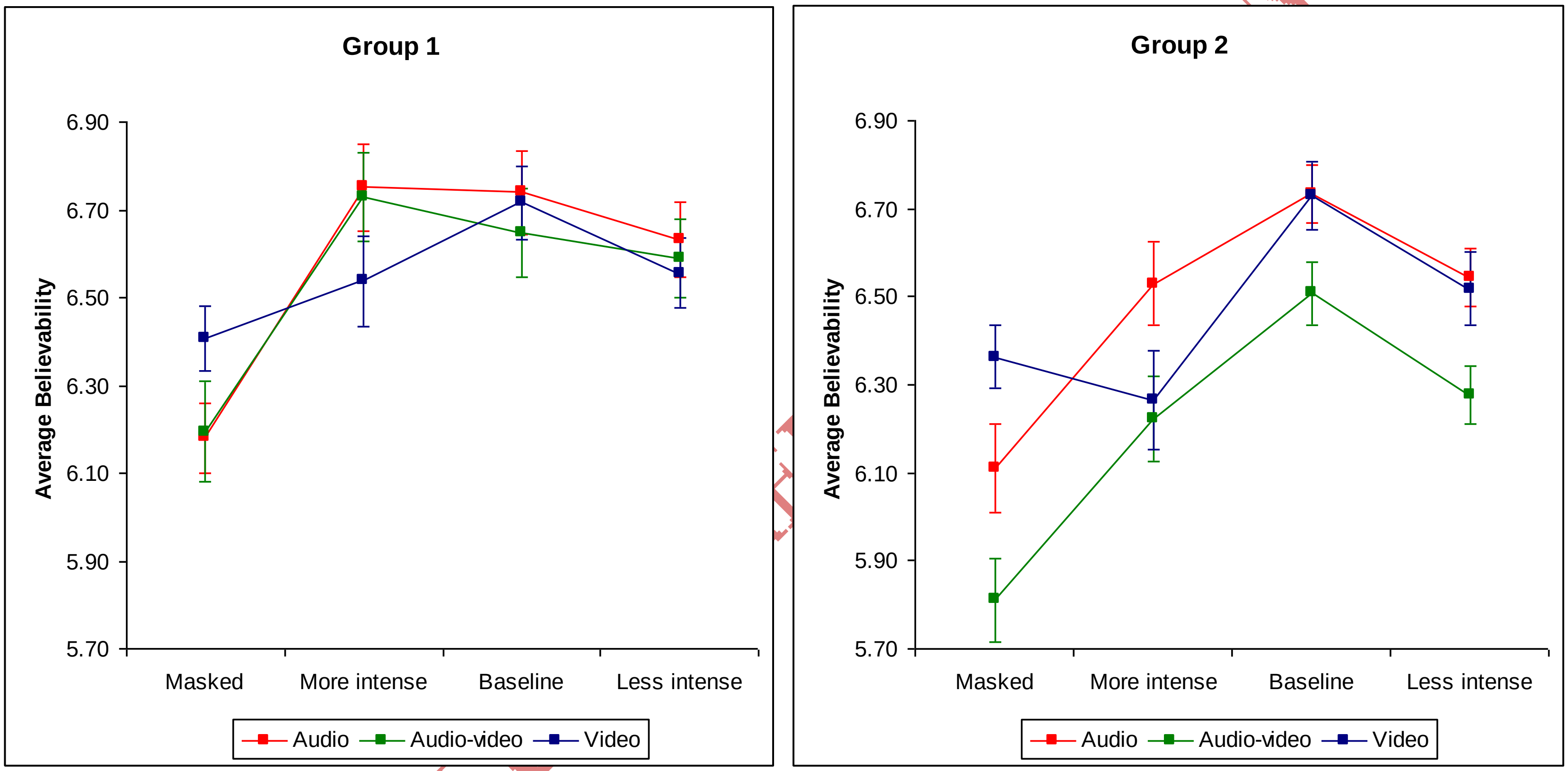

Group 1
3.80
3.60
3.40
3.20
3.00
2.80
2.60
2.40
2.20
2.00
Average Intensity
Less intense
Baseline
More intense
amu
hot
des
anx
int
ple
Group 2
3.80
3.60
3.40
3.20
3.00
2.80
2.60
2.40
2.20
2.00
Average Intensity
Less intense
Baseline
More intense
pri
irr
joy
pan
rel
sad

Figure 7


Group 1
Average Believability
6.90
6.70
6.50
6.30
6.10
5.90
5.70
Masked
More intense
Baseline
Less intense
Audio
Audio-video
Video
Group 2
Average Believability
6.90
6.70
6.50
6.30
6.10
5.90
5.70
Masked
More intense
Baseline
Less intense
Audio
Audio-video
Video