GEMEP - GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions

Tanja Bänziger
University of Geneva
40, Boulevard du Pont-d'Arve,
1205 Geneva, Switzerland
Tanja.Banziger@pse.unige.ch

Hannes Pirker
Austrian Research Inst.
for Artificial Intelligence
Freyung 6, A-1010 Vienna, Austria
Hannes.Pirker@ofai.at

Klaus R. Scherer
University of Geneva
40, Boulevard du Pont-d'Arve,
1205 Geneva, Switzerland
Klaus.Scherer@pse.unige.ch

ABSTRACT

This paper introduces the GEMEP (Geneva Multimodal Emotion Portrayals) corpus, a new repository of portrayed emotional expressions. Corpora of acted portrayals tend to include portrayals of very intense emotions, which are considered to occur infrequently in daily interactions between humans or between humans and machines. Acted portrayals have therefore been challenged as unsuited for applied research purposes. Taking a different stance, we argue that: (a) portrayals produced by appropriately instructed actors are analogue to expressions that do occur in selected real-life contexts; (b) acted portrayals – as opposed to induced or real-life sampled emotional expressions – display the most expressive variability and therefore constitute excellent material for the systematic study of nonverbal communication of emotions. We describe the guidelines used to record the corpus and some of our short-term research plans with the corpus.

Keywords

Emotions, actors, acted portrayals, multimodality, facial expressions, vocal expressions, gestures

INTRODUCTION

Most studies in the field of nonverbal communication of emotion have been performed either with acted or with posed emotional expressions (called *emotion portrayals* in this paper). Classical studies of facial expressions used photographs of people displaying selected "facial action units" (activations of single muscles or groups of muscles) and scored them for emotional expression [5]. For vocal expressions, Juslin and Laukka [3] reported that 87% of the studies they reviewed used acted portrayals. But, despite their over-representation in this research field, emotion portrayals were also always criticized for not being representative of "natural" emotional expressions [2]. Why then did several generations of researchers persist in studying them?

The most usual answer to this question lies in a number of limitations of the two alternatives to portrayals: expressions recorded under controlled conditions created to induce emotions; and expressions recorded in sampled real-life situations [6]. For practical and ethical reasons, emotions induced in experimental contexts are seldom very intense and, correspondingly, expressions in these contexts are scarce and, when they occur, faint. Expressions sampled in real-life occur in specific contexts which often cannot be reproduced, their verbal content and the overall quality of the recordings cannot be controlled, one person is usually recorded in only one or very few different emotional states, and it is not always clear on which grounds and how these states should be labelled. Hence, emotion portrayals offer a number of advantages which are difficult to obtain with either induced expressions or real-life sampled expressions: the possibility to record multiple emotional expressions produced by the same senders (actors); the possibility to obtain strong emotional expressions; and controlled recording conditions, including: uniform lexical content in speech, good acoustic and visual quality, and a clear definition of the expressive intention of the sender. Keeping identity of the sender or verbal content constant across emotions is essential to allow comparison between emotional expressions. Nevertheless, it can still be objected that this sort of control is not desirable if it is detrimental to the object of study, in this case the "natural" emotional expression and/or its context of occurrence. In the following some arguments that have been raised against the usage of emotion portrayals are discussed in more detail.

OBJECTIONS TO EMOTION PORTRAYALS

There are several common objections to the use of emotion portrayals. The classification we propose here is arbitrary but we believe that it accurately describes the main concerns that have been voiced, or may be voiced, regarding the use of portrayals in research.

Portrayals reflect stereotypes, not genuine emotions

Emotional expressions produced by actors reflect cultural stereotypes – maybe even stereotypes pertaining to specific acting schools or acting traditions. A more or less explicit assumption is that portrayals are exaggerated (over-acted) as compared to spontaneous expressions. They convey the expressive intention of the actor but not in a "realistic" way. The receiver will recognize the intention but will not necessarily believe that the sender is genuinely emotional. In cases where actors attempt to display

expressive signs without invoking emotional feelings, portrayals might indeed be produced in the absence of genuine emotions. An assumption in this view is that actors who do this might be unable to mimic (simulate or fake) the more subtle signs habitually related to the emotion. In this perspective, without being necessarily exaggerated, the portrayals would lack essential features of genuine expressions.

Portrayals represent infrequent emotions

Actors have been traditionally requested to portray so-called "basic emotions", identified by labels such as: 'anger', 'fear', 'sadness', 'happiness', or 'disgust'. Taking a closer look at the literature it becomes apparent that most studies in which acted material was used were set to study rather "extreme" emotional states, which might be better labelled: 'rage', 'panic', 'depression', 'elation', or 'repulsion'. This range of very intense emotions is considered by some authors to be unrepresentative of the range of emotions that are likely to occur on a daily basis in ordinary interactions [1]. This further led to the conclusion that portrayals are not representative of the range of emotional states that are of interest in specific applied contexts and therefore unsuited for research in corresponding fields (such as human-machine interactions).

Portrayals are decontextualized

Portrayals are produced under conditions designed to remove contextual information, in order to record variations that can be related exclusively to the emotions portrayed. Senders are recorded under identical conditions while they portray a range of emotional states. They are typically seated in front of a uniform background, facing a (video) camera. In instances were vocal or dynamic facial expressions are recorded, the actors are usually requested to first appear inexpressive, then to produce an emotional expression, and finally to get back to their inexpressive "baseline". If vocal expressions are of interest to the researchers, the senders will usually be requested to portray emotions while pronouncing the same sentence in all conditions. While most researchers have favoured this type of controlled recording design, it is certainly true that receivers presented with a relatively large number of such portrayals will not be inclined to perceive them as "natural".

ARGUMENTS FOR EMOTION PORTRAYALS

Acted portrayals allow to record strong emotional expressions and allow comparisons across emotions without systematic confounds, while the same is more difficult to achieve respectively with induced expressions and "naturally" occurring expressions. In the following paragraphs we address the objections raised in the previous section.

Portrayals need not be stereotypical nor faked

Actors can and should be encouraged to produce believable expressions by using acting techniques that are thought to stir genuine emotions through action [4]. When portraying emotions, actors should not exaggerate or fake expressions but rather attempt to reactivate emotional experiences while and through acting.
In everyday life, emotional expressions are directed to receivers with different degrees of intentionality. Some expressions might be truly "spontaneous", not directed or intentionally regulated to have an impact on a receiver, whereas acted portrayals are by definition produced intentionally and directed to a receiver. Processes underlying intentional regulation of emotional expressions and their actual effects on expressions are not well known. It would be worthwhile to systematically investigate the similarities and differences of emotional expressions produced more or less intentionally (in everyday life and/or in the laboratory). This could involve comparing acted portrayals with less "controlled" expressions, recorded under conditions that would not promote emphasis or suppression of expressions for the benefit of a receiver.

Portrayals need not to represent basic emotions only

While it is necessary to use descriptors to distinguish and conceptually organize emotional states, descriptors corresponding to "basic emotions" are probably too broad to offer a useful classification for the study of emotional expressions. For vocal expressions in particular, it has been suggested that the failure to reproduce acoustic profiles reported for a given category – for example 'fear' – across studies might derive form variations on the level of the definition of that category [6]. In different studies emotions labeled 'fear' might correspond to very different states ranging from 'apprehension' to 'panic', which obviously would result in different expressive patterns.

Acted portrayals should clearly not be restricted to "basic emotions". Actors can be requested to portray states that are of interest in specific research contexts (for example "frustration" in response to a malfunctioning device or computer application). Researchers need to have clear aims and correspondingly clear operational definitions of the emotional states of interest in their field. In interactions between humans and computers, very intense reactions might be triggered (for instance when playing a challenging game), but more subtle reactions might be of importance as well (e.g. "frustration"). Portrayals produced by appropriately instructed actors could reflect a variety of "realistic" states in different contexts, the selection and definition of the states portrayed pertains to the researcher.

Add more "context" to emotion portrayals

In most cases, emotional reactions are tightly linked to a specific context of occurrence (a situation/event that triggers the reaction). Portrayals would probably appear more "natural" if they were not produced in the

complete absence of eliciting events. In everyday life, it is quite rare that a calm, inexpressive person suddenly becomes very emotional, without any apparent reason, and within a few seconds recomposes herself, appearing perfectly calm and inexpressive again. Furthermore, it is undoubtedly beneficial to formulate precise operational definitions of the emotions to be portrayed. This should include at least a rough description of the situation/event in which the reaction takes place. Hence the minimal "context" defined for a given portrayal could be composed of a scenario describing the situation in which the emotion is elicited and – as it is probably more easy for an actor to direct the portrayal at a designated receiver – a brief interpersonal interaction taking place in this situation.
For the construction of any corpus featuring emotional expressions it is essential to clearly define the range of emotions included. Many objections raised against emotion portrayals would probably lapse if the portrayed states would be better selected and clearly defined. For instance, the notion that acted portrayals are lacking "naturalness" (are stereotyped or exaggerated) might be largely derived from the insistence on recording portrayals reflecting extreme emotions.

PROCEDURE USED FOR RECORDING THE CORPUS

This section outlines some of the guidelines applied to record the GEMEP, with an emphasis on the aspects already introduced in the previous sections.

Selection of portrayed emotions

The affective states portrayed were partly selected to match the states frequently studied in the literature dealing with facial and/or vocal expressions of affect. Some less frequently examined states were also included in order to address specific research questions. A relatively large number of positive states – such as 'pride', 'amusement', 'elation', 'interest', 'pleasure', or 'relief' – was for example included in order to challenge the traditional view according to which only one rather 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 fulfils at least two aims: (1) Reviews of studies describing acoustic profiles of emotional expressions have recurrently reported differences in acoustic features of vocal expressions mostly related to arousal level, the crossing of arousal level and emotion category should allow to partly 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 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'.

A further attempt to increase the variability of the expressions was undertaken by requesting the actors to produce some of the emotions with less intensity and with more intensity than the intensity that they thought corresponded to the most 'usual' intensity for a given emotion. An underlying assumption (which remains to be tested) is that the portrayals produced with less intensity might be closer to expressions that could occur in daily interactions, while 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 request to partially mask some of the expressions (i.e. to portray a relatively unsuccessful deception attempt for some of the affective states).

Definition of emotions and of their "context"

Short definitions of the emotional states and "scenarios"[1] were provided to the actors several weeks before the recordings took place. Three "scenarios" were created in order to instantiate each affective state. 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 while pronouncing two pseudo-linguistic sentences (1. "ne kal ibam sud molen!"; 2. "kun se mina lod belam?"). The actors were further requested to express each affective state while uttering a sustained vowel, which allowed recording brief emotional expressions in the absence of articulatory movements. For each affective state, the director and the actors were trying different "scenarios" and – after a period of "rehearsal" – recorded one or more interactions until they were satisfied with their performance.

Technical aspects and description of the corpus

Ten professional French-speaking actors portrayed 15 affective states under the direction of (and in interaction with) a professional stage director. Three digital cameras were used for simultaneously recording: (a) facial expressions and head orientations of the actors, (b) body postures and gestures from the perspective of an interlocutor, (c) body postures and gestures from the perspective of an observer standing to the right of the actors (cf. Fig. 1). Sound was recorded using a separate microphone at each of the three cameras, plus an additional microphone positioned over the left ear of the actor, providing a separate speech recording with a constant distance to the actor's mouth.

[1] Definitions and scenarios are currently available only in French and can be obtained on request to the first author.


a

b

c

Figure 1: still frames illustrating the 3 camera angles used in the video recordings

Video and audio data was 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 in separate files. Over 7'300 such sequences, among them about 5'000 containing the pseudo-linguistic sentences and the sustained vowel, were extracted from the original interactions. However the corpus will be reduced to less than 2'000 sequences by removing redundant sequences; 2-3 repetitions will be selected for every expected condition based on expert ratings.

The selected sequences will be thoroughly documented in terms of lay ratings of emotional expressivity, "naturalness" of expressivity, and accuracy of emotional communication. The audio recordings are currently automatically segmented, i.e. phoneme and syllable boundaries are determined, using phonetic alignment methods.

RESEARCH PLANS AND POTENTIAL APPLICATIONS

The uniform structure of the recordings makes this corpus especially well suited for systematic quantitative analysis and comparisons across emotions. This holds true for both the acoustic as well as the visual aspects of the data. The first results of the phonetic alignment and segmentation of the standard sentences are promising. The prospective availability of fine-grained phonetic segmentations will allow for phonetic studies that go well beyond the measurement of global prosodic parameters. For example durational effects can be measured at the level of single phonemes. In addition an exact phonetic transcription makes investigations on pausing and articulatory precision possible. Special attention will be paid to the investigation of voice quality parameters. The availability of robust methods for measuring these parameters is still an issue. But due to its size and diversity the corpus can also be used as either training- or test-data for feature-extraction and classification methods and thus can contribute to the improvement of these methods.

In the visual domain, first trials with hand- and face-tracking are performed. One aim is to prepare the ground for subsequent studies of the effect of emotions on the temporal and spatial properties of hand-gestures, in the sense of Wallbott [7] but with the possibility to use truly quantitative measurements on a much broader data set. Subsets of facial expressions will be manually FACS coded. In a further step, manual FACS annotations could be compared to, and potentially supplemented with, automatic extraction methods applied to facial movements. More importantly, we hope to be able to extract data that will allow comparisons across modalities (voice, gestures, face) for different emotions. One issue in this respect would be to assess the extent of "synchronisation" between modalities for various emotional expressions.

The multimodal nature of the corpus will also allow investigating the effects of the different modalities on recognition accuracy for different emotions in rating tests. It will be possible to evaluate the performance of both human subjects and affect-recognition systems when presented with (unimodal) vocal, facial or gestural/postural expressions in comparison to accuracy for combinations of these modalities.

The selection of states portrayed was essentially driven by research questions and applications stemming from emotion psychology. The recordings should in particular allow to investigate questions related to the processing of emotional faces, voices and gestures in neuro-imaging studies and to develop assessment tools of multimodal emotional sensitivity for normal and clinical populations.

CONCLUSIONS

Acted portrayals can provide very valuable contributions to the study of multimodal expressions of emotion. The emotional states portrayed should be carefully selected and defined according to specific research questions. Nevertheless comparison with expressions recorded under more "spontaneous" conditions remain necessary to qualify the results obtained with portrayals.

Acknowledgements

This research is supported by the Swiss National Science Foundation (FNRS 101411-100367), the EU Network of Excellence HUMAINE (IST 507422) and by the Austrian Funds for Research and Technology Promotion for Industry (FFF 808818/2970 KA/SA). OFAI is supported by the Austrian Federal Ministry for Education, Science and Culture and by the Austrian Federal Ministry for Transport, Innovation and Technology.
This publication reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

REFERENCES

1. Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. Speech Communication, 40, 5-32.
2. Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. Speech Communication, 40, 33-60.

3. Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin, 129, 770-814.
4. Moore, S. (1984). The Stanislavski system. The professional training of an actor. New York: Penguin Books.
5. Russell, J., Bachorowski, J., & Fernandez-Dols, J.-M. (2003). Facial and vocal expressions of emotions. Annual Review of Psychology, 54, 329-349.
6. Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication, 40, 227-256.
7. Wallbott, H. G. (1998). Bodily expression of emotion. European Journal of Social Psychology, 28, 879-896.