Coding Emotional Events in Audiovisual Corpora

L. Devillers, J-C. Martin

LIMSI-CNRS; France
{devil, martin}@limsi.fr

Abstract

The modelling of realistic emotional behaviour is needed for various applications in multimodal human-machine interaction such as the design of emotional conversational agents (Martin et al., 2005) or of emotional detection systems (Devillers and Vidrascu, 2007). Yet, building such models requires appropriate definition of various levels for representing the emotions themselves but also some contextual information such as the events that elicit these emotions. This paper presents a coding scheme that has been defined following annotations of a corpus of TV interviews (EmoTV). Deciding which events triggered or may trigger which emotion is a challenge for building efficient emotion eliciting protocols. In this paper, we present the protocol that we defined for collecting another corpus of spontaneous human-human interactions recorded in laboratory conditions (EmoTaboo). We discuss the events that we designed for eliciting emotions. Part of this scheme for coding emotional event is being included in the specifications that are currently defined by a working group of the W3C (the W3C Emotion Incubator Working group). This group is investigating the feasibility of working towards a standard representation of emotions and related states in technological contexts.

1. Introduction

Interest in affective computing has grown over the past decade. The success of the recent international conference on "Affective Computing and Intelligent Interaction", and the developments from the European Network of Excellence Humaine (www.emotion-research.net), show that "emotions" are now considered as significant factors in the design of Human-Computer Systems. The choice of the appropriate corpus for training computational models is fundamental. The training data must be as close as possible to the behaviours observed in the real application. Thus, expressions of emotion should be collected as they occur in everyday action and interaction rather than as idealised archetypes.

Natural databases highlighted a descriptive challenge: the expressions of emotions that they contain are not adequately described by a few theoretically derived labels, such as basic emotions (Cowie and Cornelius 2003, Devillers, Vidrascu & Lamel 2005, Devillers et al., 2006). There is indeed a significant gap between the affective states observed in artificial data (acted data or induced data) and those observed with "real-life spontaneous data in situ". This difference is mainly due to the "Real" context. We define "context" here as the events that are at the origin of the affective state of a person.

In our previous empirical studies on real-life audio and audio-visual data (Devillers & al, 2005) (Devillers, Abrilian, Martin, 2005), we have shown that there are many complex mixtures of emotion. In the same idea, (Wilhelm, Schoebi & Perrez 2004) have shown that while people never report their state as being completely unemotional, examples of full-blown emergent emotions are really quite rare.

Appraisal is the process by which an organism assesses its overall relationship with its environment including not only its current condition but past events that led to this state as well as future prospects. In this paper, we describe a study exploring how to code the events that triggered spontaneous expressions of real-life emotions in audiovisual data.

Cognitive appraisal theory argues that an organism may possess many distributed processes for interpreting this relationship (e.g., planning, explanation, perception, etc.) but that appraisal maps characteristics of these disparate processes into a common set of intermediate terms (intermediate between stimuli and response, between organism and environment) called 'appraisal variables' (Gratch and Marsella, 2004). Scherer's appraisal dimensions (Sherer et al., 2001) have been studied in emotion recall experiments and have been used to predict major modal emotions. We already used the appraisal dimensions in emotion perception investigation for describing complex blended emotional behaviour of the video-taped person (Devillers et al. 2006). This previous study also pointed out some difficulties in the manual labelling of appraisal dimensions. One of the main difficulties with the appraisal categories is that they are inherently linked to a focal event or situation. In our studies, the naturalistic clips often imply that the emotional behaviour relates to multiple events (e.g. the videotaped person is reacting to multiple events: a trauma which happened several years ago, a recent illness, and the current interview). Parts of the appraisal framework cannot be applied without establishing which of these events is to be annotated. For several appraisal dimensions and the associated items, the existence of multiple relevant events is key; it has an impact on the following dimensions: agent responsible, motive, outcome probability, and all the items of the coping potential dimension except power - controllability of immediate event, controllability of consequence event, and possible adjustment to person's own goals. The same problem affects the 'compatibility with standards'.

It is also meaningful for complex emotional labels such as masked or blended emotions to identify the existence of multiple relevant events. Different types of mixtures of emotions are considered: simultaneous emotions experienced due to the presence of several emotional events (such as having fear and being angry at the same time, but for two different reasons); and regulation (such as trying to mask one emotion with another one). The emotional events can be trigger, cause, or eliciting event of an emotion; and the object or target of the emotion, that is, what the emotion is "about" (Schröder et al. 2007). Trigger and target events are conceptually different; they may or may not coincide. The distinction between both events is not straightforward.

Our aim in this study is to describe the emotional events linked to these complex emotional behaviours. Are we able to infer them from the context we see (the clips are very short, sometimes few seconds)? Are we able to reliably annotate these emotional events and their temporal relationships? The main difficult point of this representation is to find the useful levels of description in term of granularity and temporality.

We have collected two audio-video corpora (Abrilian et al., 2005) (Zara et al, 2007), one is extracted from TV news (naturalistic corpus), the other is the record of a two-players game designed in order to induce emotions. The first corpus contains monologues, and the second corpus features dyadic interactions. In both corpora, complex spontaneous emotional behaviours were observed.

In section 2, we describe the scheme we propose to code perceived emotional events in audiovisual data. Section 3 reports the EmoTV corpus used for the empirical study. Then, we discuss the protocol; the strategies used and triggered events for eliciting emotions in the game EmoTaboo. Finally, we conclude on the issues of this study in section 5.

2. Perceived Emotional Events Scheme

Our group has been working for several years on the problems brought by the coding of emotional behaviors. The scheme that we have defined for the coding of emotional behavior features multiple types of descriptor – verbal labels, abstract dimensions and contextual annotations such as a plain text annotation of the global emotional situation (Devillers et al. 2005).

The aim of the current study was to consolidate our expertise on annotation by defining and evaluating a scheme for coding. The main goal was to find out the most promising descriptors for describing events. Such knowledge on the representation of emotional events might be useful for annotating videos of emotional behaviors collected during human-human interactions, during human-computer interaction, and also for defining protocols for eliciting emotion thanks to the induction of appropriate events.

Different events can trigger different affective states: emotions / attitudes / interpersonal stances at the same time. As an example, we can imagine a physical internal event such as "a stomach-ache" that triggers pain and an external event as "someone helping the sick person" that triggers relief.

The definition of an event in our study is that the event is:
- perceived in a video clip as triggering the observed emotional behavior
- described with neutral words
- and consensual for 3 annotators

Being interviewed and videotaped itself is a common event to all clips. For all the other emotional events, a list of events assessed as being relevant to our set of clips, was defined. Three groups of emotional events are defined in the OCC model (Ortony, Clore, & Collins, 1988) depending on what is evaluated: the consequences of events for oneself or for others, the actions of others and the perception of objects. In our corpus, we only observed events that are being evaluated with respect to their consequence for oneself or others and actions of others. For example, in one of our clips (#ext41), a lady is reacting to the consequences of elections in which her political party lost for herself and for her political party. In another clip (#ext3), a lady is accusing other people (she is reacting to their actions). We have also annotated longer events such as "a trial which is going on", "election campaign", "football match". These long-term events are more connected with long-term affective states such as mood or positive or negative attitude.

Our events scheme permits to annotate up to 3 emotional events for a given clip. Each event is annotated according to the following temporal dimensions:

Temporality

- *Past (> 1 week)*
- *Near past (< 1 week)*
- *Present (today)*
- *Near future (< 1 week)*
- *Future (> 1 week)*

Duration

- *Months or more*
- *Weeks*
- *Days*
- *Hours*
- *Minutes*

Annotators can select up to three events, and have to specify the relations between these events. The new idea of this scheme is to add a temporal aspect of the events. We first classify events in short-term and long-term events. Then we specify the temporal-relation between events.

3. The EmoTV study

3.1. The corpus

The EmoTV corpus is described in details in (Devillers, Martin, Abrilian, 2005). It is a collection of video-clips, mainly extracted from TV news.

3.2. Theoretically derived descriptions

Theory suggests many possible ways for describing emotions other than everyday labels. The scheme that we defined here incorporates what seems to be the most important options for coding the events that we observed in our set of video clips.

A strong tradition distinguishes emotions in terms of the 'appraisals' that they involve. Appraisals are perceptual evaluations of emotion-relevant aspects of the situation on a set of dimensions suggested by Scherer and his colleagues (Sander et al 2005). The following set of labels was used to describe the protagonist's appraisal of one event E at the focus of his/her emotional state. These items are grouped under four broader headings.

- *Relevance* Suddenness of E; Familiarity of E; Predictability of E; Intrinsic pleasantness of E; Desirability of the consequences
- *Implications* Agency responsible (self / other / group / nature); Underlying motive (negligence / intent); Nature of likely consequences (negative vs. positive); Relation to expectation (consonant to dissonant); Conduciveness to goals (conducive vs. obstructive); Urgency (low vs. high)
- *Coping potential*; Controllability (low vs. high); Controllability of consequences (low vs high); Power of person to change the outcome of these events (low vs. high); Scope for person to adjust own goals (low vs. high);
- *Compatibility of the situation with standards* With external standards (norms or demands of a reference group); With internal standards (self ideal or internalized moral code)

These are brief descriptions – fuller explanations are given by Sander et al. (2005).

3.3. Everyday labels that apply to naturalistic samples

Emotions in a strict sense	Emotion-related states	
Anger (hot)	Affection	Interest
Anger (cold)	Amusement	Irritation
Contempt	Anxiety	Pleased
Despair	Boredom	Relief
Disgust	Courage	Relaxation
Elation/joy	Disappointment	Satisfaction
Fear	Doubt	Serenity
Guilt	Embarrassment	Shock
Happiness	Empathy	Stress
Pride	Excitement	Worry
Sadness	Friendliness	
Shame	Helplessness	
Surprise	Hope	

Table 1: everyday labels selected for the study

We use the term 'everyday labels' (to describe words of the kind listed in Table 1 above. Table 1 aggregates the labels that were observed as being relevant to the annotation of naturalistic TV clips material (Devillers et al., 2006).

3.4. Events annotation

We defined a first scheme of annotation and tested it on a subset of EmoTV (28 clips). Our annotation strategy is to proceed in two steps. First we identify and annotate the emotional events. Second, we link the annotation of emotions to each of the events.

Video	Events	Emotions
Ext02 	**1-Lawsuit (current)** : Temporality: past Duration : months and more **2- Someone is accusing other people** Temporality: present Duration : hour **3- The trial has just finished** Temporality: present (D-day) Duration : minutes	Anger Despair Disappointment Disgust Helplessness Worry

Ext03 	**1 - Lawsuit and confinement of her father and brother** Temporality: past Duration : months and more **2- The trial has just finished** Temporality: present (D-day) Duration : minutes	Anger Despair Disgust
Ext22 	**1- Bathing in the see** Temporality: present (D-day) Duration : hours 2- **Examination of the French school diploma BAC in one month** Temporality: future Duration : days	Serenity Worry
Ext24 	**1- Football cup** Temporality: past Duration : months and more **2- Football match** Temporality: close future Duration : hours	Serenity Pride
Ext36 	**1- Election campaign** Temporality: past Duration : months and more **2-Election results in the GQ** Temporality: present (D-day) Duration : minutes	Anger Sadness Shame Disgust
Ext41 	**1- Election campaign** Temporality: past Duration : months and more **2- Election results** Temporality: present (D-day) Duration : minutes	Courage Disappointment Joy
Ext93 	**1- Problem of environment** Temporality: past, present, future Duration : months and more	Irritation
Ext82 	**1- Swindle in the "batiment"** Temporality: past, close past Duration : months and more	Irritation
Ext97 	1-**The increase of the price of the coffee** Temporality: present (D-day) Duration : minutes	Irritation

Figure 1: examples of events annotation extracted from EmoTV

The scheme that we have defined enables s to annotate up to 3 emotional events. We use a list of pre-defined events collected after a first study of the corpus. Each event is annotated according to the following two temporal dimensions: temporality (past, close past, present, close future, future) and duration (months and more, weeks, days, hours, minutes).

The Figure 1 provides some examples of coding emotional events for 9 clips from EmoTV. We often observe two events per clip: one long-term event (ex: law-suit, election campaign) and one short-term event (ex: exit of audience, election results). Most of EmoTV clips have been annotated with several emotion labels such as Courage/Disappointment (#41) or Anger/Despair (#03). The presence of multiple events per clip is correlated with these annotations of complex emotions. For example, in the clips #82, #97, #93, only one event has been annotated and a smaller number of emotion labels was used for the annotation of emotion. In our scheme, it is possible to give for temporality a large plage of temporality including past, present and future such as in the clip about problem of environment (#93). This exploratory study showed the diversity of events that can elicit complex affective states in TV interview. In this study, we use the general term of emotion or emotion related state instead of the term of affective state. In fact, very few of the cases listed include real emotions as described in the literature: short-time and high intensity (only the case where Temporality: present (D-day), Duration: minutes). As we shown, complex emotions are often a mixture of emotion and other kind of affective states which involve different timescales (moods, attitudes…).

4. The EmoTaboo Protocol

Whereas the EmoTV corpus described in the previous section focused on spontaneous expression of emotion during monologues. In order to study the impact of dyadic interaction on the expression of emotion we defined another protocol called EmoTaboo. EmoTaboo is an adaptation of the game Taboo. It involves interactions between two players. One of them has to guess a word that the other player is describing using his own speech and gestures, without uttering five forbidden words. The word to guess and the five forbidden words are written on a card. Each person had to make guess three series of words alternating roles (mime and soothsayer). The two players do not know each other. One of them is a naïve subject whereas the other player is instructed. This confederate knew all the cards in advance, and for each card, indications were given on how to induce emotions in the naïve subject (e.g. do not find the word on purpose). We involved a confederate in the protocol because we wanted to be sure to collect enough emotional interactions and we supposed that it would enable us to have a better control over the emotion elicitation situations. To ensure the engagement of the subjects in the task, the results of previous teams were displayed on a board in the room during the game, and a gift token was promised to the winner team.

4.2 Strategies in EmoTaboo

We used strategies for eliciting emotions at three different levels in the procedure: in the course of the game, in the selection of the cards, and in the directions given to the confederate.

Strategies connected to the course of the game. The mime had ten seconds to read the card (on which was written the word to make guess and five forbidden words). Then, he had two minutes to make guess the word. Thirty seconds before the end of the prescribed time, the experimenter announced the remaining time in order to motivate the players and to elicit stress. After these two minutes, the experimenter took stock of the penalties in case the secret word was not found or if the team transgressed some game rules (e.g. using a forbidden word).

Strategies connected to the selection of the cards. Game cards were provided to the players in ascending order of difficulty. Regarding the type of this game, we supposed that the emotions induced by game cards would include embarrassment, shame, amusement and surprise. To ensure their elicitation, we played on the knowledge of the word, the easiness to guess the word, and what the word evokes in players. We chose cards containing very uncommon words (e.g. "palimpsest") supposed to arouse embarrassment or shame, words evoking disgusting things (e. g. "putrid") or words with sexual connotation (e.g. "aphrodisiac").

Strategies connected to the instructions given to the confederate. For each card, the confederate received instructions such as "do not find the word on purpose", "propose words with no relation at all with what is said by the naïve player". For each card, a list of emotions to elicit from the naïve subject was given (e.g. card "temptation": negative emotions: disappointment, frustration, stress; positive emotions: pride, satisfaction). For each emotion an illustrative list of possible strategies was proposed (e.g. to induce anger, criticize the naïve player").

We recorded ten pairs of players, each pair using twenty cards. Naïve subjects were university students (four women and six men), Confederates were close relations of the experimenter or laboratory staff (three women, five men). We collected about eight hours of videos with four different viewpoints corresponding to face close-up and upper body of both players (Figure 2).

Figure. 2. The collected data features four viewpoints. The naive subject is on the left side, the confederate is on the right side.

4.3 Events in EmoTaboo

For EmoTABOO, the timescales of the events eliciting emotions is not the same as those of the events observed in EmoTV: the EmoTABOO events are rather short term. The only long-term event is the game session. The observed events can be also described in EmoTABOO in term of dimensions of appraisal:
For example the fact that the card contains the word "palimpsest" is estimated by the player as:
- **recent**
- **unexpected** (the player did not expect a word that he did not know)
- **incompatible** with his(her) purposes to win the game.
Other events eliciting emotions include:
- The experimenter announces a penalty,
- the confederate proposes words with no relation at all with what is said by the naïve player
- the confederate finds the mimed word/does not find the mimed word
- the confederate is ironic
- the confederate criticizes the naïve player.
All these events can be described in term of appraisals and temporality. In that corpus, a lot of complex emotions has also been annotated that can be correlated to events. As an example, being relieved that the player find the mimed word and at the same time feeling disappointed because the experimenter announces a penalty.

5. Conclusion

Describing emotional events in audiovisual corpora is required:

- for building emotional eliciting protocol,
- for studying appraisals variables,
- and for explaining complex emotions.

In the exploratory study described in this paper, we tried to show the potential of a scheme of emotional events in order to better understand the emotional states. We proposed to code the temporal aspects of the events: date and duration (short-time, long-time). We also proposed to code the events using the appraisal dimensions.

Future research include 1) studying the relations between the events and the perceived complex emotions, and 2) studying the relations between temporal and appraisal annotations.

Acknowledgement

This work was partly funded by the FP6 IST HUMAINE Network of Excellence (http://emotion-research.net).

6. References

Abrilian, S., Martin, J-C., Devillers, L. (2005). A Corpus-Based Approach for the Modeling of Multimodal Emotional Behaviors for the Specification of Embodied Agents. *HCI International 2005*, Las Vegas, USA.

Cowie & Cornelius (2003). Describing the emotional states expressed in speech. *Speech Communication*, 40(1-2), pp. 5-32.

Devillers, L., Abrilian, S. and Martin, J.-C. (2005) Representing real life emotions in audiovisual data with non basic emotional patterns and context features. *1st International Conference on Affective Computing & Intelligent Interaction (ACII'2005)* Beijing, China, October 22-24.

Devillers, Vidrascu & Lamel, (2005). "Challenges in real-life emotion annotation and machine learning based detection", *Neural Networks* 18, pp. 407-422.

Devillers, L., Vidrascu, L. (2007), Emotion recognition, *«Speaker characterization », Christian Müller, Susanne Schötz (eds.), Springer-Verlag.*

Devillers, L., Cowie, R., Martin, J.-C., Douglas-Cowie, E., Abrilian, **S.**, McRorie, M. (2006) Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. *(LREC 2006)*, Genoa, Italy, 24-27 may.

Gratch, J., Marsella, S., (2004) A domain independent framework for modeling emotion. Journal of *Cognitive Systems Research*, 5 (4), 269-306.

Martin, J.-C., Abrilian**, S.**, Devillers, L., Lamolle, M., Mancini, M. and Pelachaud, C. Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. *5th International Working Conference On Intelligent Virtual Agents (IVA'2005)* Kos, Greece, Sept 12-14.

Ortony, A., Clore, G.L., Collins, A. (1988). *The cognitive structure of emotions.* Cambridge, U.K., Cambridge University Press.

Sander, D., Grandjean, D., & Scherer, K. (2005) A sytsems approach to appraisal mechanisms in emotion *Neural Networks* 18, pp 317-352.

Scherer, K. R., Schorr, A. and Johnstone, T., eds. (2001). Appraisal Processes in Emotion. *New York: Oxford University Press*.

Schroeder, M., Devillers, L., Karpouzis, K., Martin, J-C., Pelachaud, C., Peter, Ch., Pirker, H., Schuller, B., Tao, J. and Wilson I. (2007): What should a generic emotion markup language be able to represent?, *ACII07, 2nd International Conference on Affective Computing & Intelligent Interaction*.

Wilhelm, P., Schoebi, D., Perrez, M. (2004). Frequency estimates of emotions in everyday life from a diary method's perspective: a comment on Scherer et al.'s survey-study "Emotions in everyday life". *Social Science Information* 43: 647-665.

Zara, A., Maffiolo, V., Martin, J-C., Devillers L. (2007): Collection and Annotation of a Corpus of Human-Human Multimodal Interactions: Emotion and Others Anthropomorphic Characteristics, *ACII07, 2nd International Conference on Affective Computing & Intelligent Interaction*.