EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces

Abrilian, S., Devillers, L., Buisine, S., Martin, J.-C.

LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France
{sarkis, devil, buisine, martin}@limsi.fr

Abstract

The development of future multimodal affective interfaces such as believable Embodied Conversational Agents requires to model relations between natural emotions and multimodal behaviors in various real-life contexts. Research should thus explore multimodal corpora and real-life databases (e.g. news, natural conversations). In order to collect reliable emotion annotations, one also needs to understand how people perceive emotional expressions in audio channels (vocal cues, content, and grammatical issues), and in visual channels (facial expressions, body movements, gaze, and spatial behavior). We designed EmoTV1, a corpus of video clips recorded from French TV channels containing interviews. In order to study the influence of the modalities on the perception of emotions, two annotators used the Anvil tool to annotate this corpus for three conditions: 1) audio without video, 2) video without audio, 3) video with audio. The annotators segmented the emotions perceived. Their classification resulted in a set of 14 labels: anger, despair, disgust, doubt, exaltation, fear, irritation, joy, neutral, pain, sadness, serenity, surprise and worry. The annotators also used two classical appraisal dimensions: intensity and valence (negative/positive emotion). Our work, described in this paper, lists the advantages and drawbacks of the corpus, describes how to gather a relevant corpus of video clips and define a scheme for annotating emotions. It also describes the possible annotation disagreements and the improvements that we intend to make after the analyses of this first annotation phase.

1 Introduction

Multimodal Human-Computer Interfaces aim at enabling the combined use of several communication modalities between the user and the computer. Amongst them, Embodied Conversational Agents (ECAs) make use of a wide range of "natural" modalities such as speech, gesture, and facial expressions (Cassell, Sullivan, Prevost & Churchill, 2000 ; Paiva, 2000) but require to model relations between natural emotions and multimodal behaviors in various real-life contexts. Research should thus explore multimodal corpora (Wegener Knudsen, Martin, Dybkjær, Berman, Bernsen et al., 2002 ; Martin, den Os, Kuhnlein, Boves, Paggio et al., 2004) and real-life emotional video databases (e.g. news, natural conversations) (Douglas-Cowie, Cowie & Schröder, 2000). In order to collect reliable emotion annotations, one also needs to understand how people perceive the emotional expressions in audio channels (vocal cues and linguistic issues), and in visual channels (facial expressions, body movements, gaze, and spatial behavior).

Three types of emotion annotation are generally used in research on emotion: appraisal dimensions, abstract dimensions and most commonly verbal categories. These verbal categories include both “primary” labels (anger, fear, joy, sadness, etc. (Ekman, 1999) and “secondary” labels for social emotions (e.g. love, submission). (Plutchik, 1994) combined primary emotions to produce other labels for “secondary” emotions. For example, love is a combination of joy and acceptance, whereas submission is a combination of acceptance and fear. Yet, the number of labels required for annotating real-life emotions might be very high when compared to basic emotions. Actually, most of the emotions modeling studies have used a minimal set of labels to be tractable (Batliner, Fisher, Huber, Spilker & Noth, 2000 ; Devillers & Vasilescu, 2004). Instead of using these limited numbers of categories, some researchers define emotions using continuous abstract dimensions: Activation-Evaluation (Douglas-Cowie, Campbell N., Cowie R. & Roach P., 2003), Intensity-Evaluation (Craggs & Wood, 2004). But, these dimensions do not allow precise emotion representation, for example it is impossible to distinguish between Fear and Anger. Finally, the appraisal model is useful for describing the perception / production of emotion. The major advance in this theory is the detailed specification of appraisal dimensions that are assumed to be used in evaluating emotion-antecedent events (pleasantness, novelty, etc) (Scherer, 2000).

The work presented in this paper deals with theoretical issues, for the study of naturalistic and non-acted data, exploring how to annotate real-life non-basic emotions and how to define a typology of non-basic emotions. Section 2 describes the emotional corpus collected. Section 3 details the context annotation phase. Section 4 presents a first emotion annotation phase: segmentation and labeling. Section 5 details the analysis of the annotations results and discusses them. Section 6 presents our future works on the relations between emotion and multimodal behavior and on the specification of real-life emotions in ECAs, more particularly on the specification of mixed emotions.

2 The EmoTV1 corpus

Our goal is the study of multimodal behaviors and real life emotions. We used the following criteria for collecting our emotional video corpus: TV interviews (monologue), realistic situations, presence of emotion, visibility of speaker's face and upper body, multimodal signs (speech, head, face, gaze, gesture, torso), and French language. This resulted in a corpus of 51 video clips recorded from French TV channels (Figure 1), interviews from news on 24 different topics (Figure 2). 48 different persons were interviewed, in a wide range of positive/negative emotions.

Figure 1: Samples of interview contained in EmoTV1

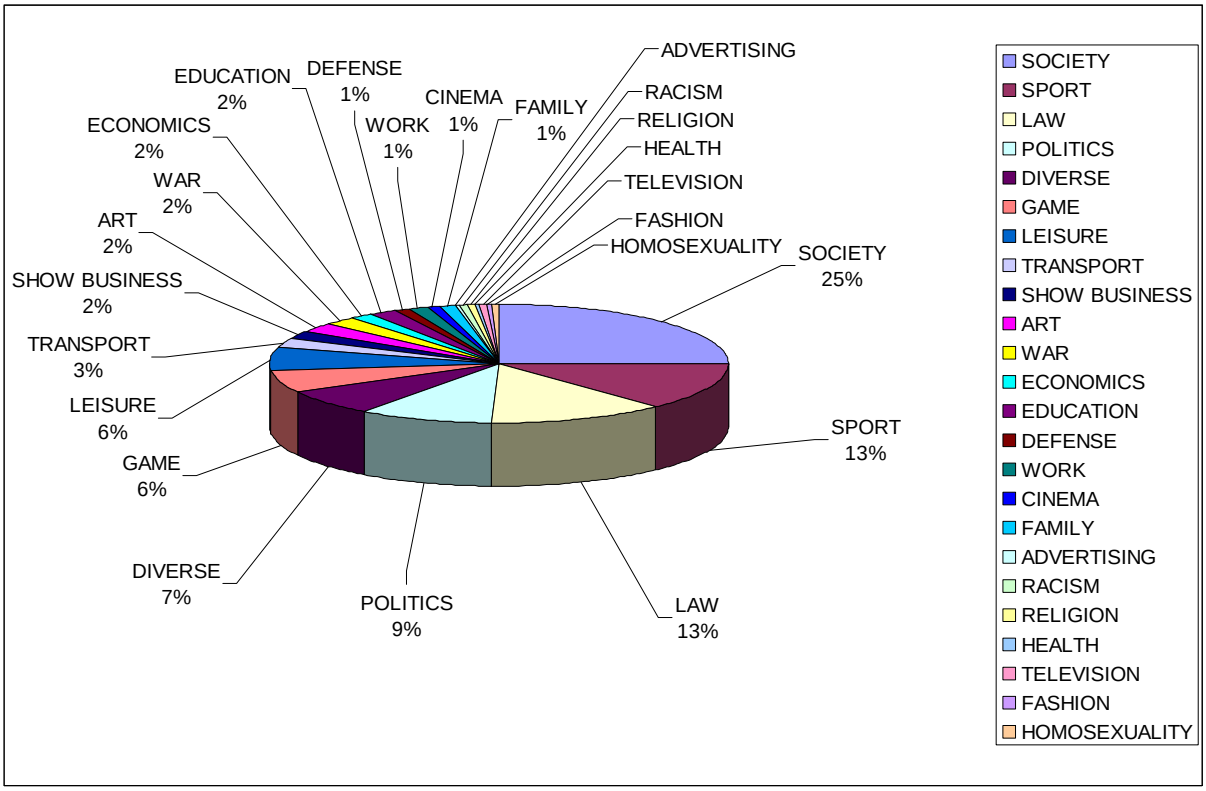

Figure 2: Topic distribution of EmoTV1

The total duration of the corpus is 12 minutes; clips range from 4 to 43 seconds, the speech transcription contains a total of 2500 words, of which 800 distinct words. These videos show the complexity of real-life mixed emotions comparing to acted emotion corpora.

The corpus has some advantages, as the presence of spontaneous emotional behaviors, a wide variety of contexts, and the illustration of requirements on annotation schemes at several levels. It also has some drawbacks due to TV video interviews, such as the eventual lack of visibility of facial expressions due to glasses, hairs, or beard, the lack of visibility of some gestures, and the low quality of video. Nevertheless, it fits our goal since we are interested in the study of spontaneous instead of acted emotions and we do not intend to do automatic processing of the video cues.

This corpus has been annotated by two coders at several levels using a context and emotion coding scheme that we describe in the next sections.

3 Context annotation

We have selected the attributes (Table 1) for describing contextual information as relevant for the study of emotions.

Table 1: Attributes used for coding context

Context attribute	Description
Theme	Topic of the interview
Degree of implication	Estimation of the level of implication of the speaker
To whom	Indicates towards whom the emotional behavior is directed
What for	What is the communicative goal perceived in the whole video
Causes of emotion	Perceived causes of emotion

The “theme” attribute consists in a free text field, annotated by each coders, and after contextual annotation, the normalization of the 100 written themes resulted in 24 topics.
The “degree of implication” is estimated with value between 1 (very low implication) and 5 (very high implication).
The “to-whom” attribute is coded using a free text field.
The “what for” attribute (free text) corresponds to the global communicative act. After normalization, 22 different communicative acts were listed (Table 2).

Table 2: The values (communicative acts) of the “What for” attribute

Communicative act	Communicative act
To complain	To transmit
To criticize	To teach
To revolt	To communicate
To blame	To testify
To claim	To explain
To demonstrate	To convince
To describe	To justify
To exhibit	To encourage
To express	To reassure
To show	To joke
To share	To divert

Finally, the attribute “Causes of emotion” could take one of the two values: external circumstances, or self behavior.

The next picture (Figure 3) shows a sample of emotion context annotation with the Anvil annotation tool (Kipp, 2001):

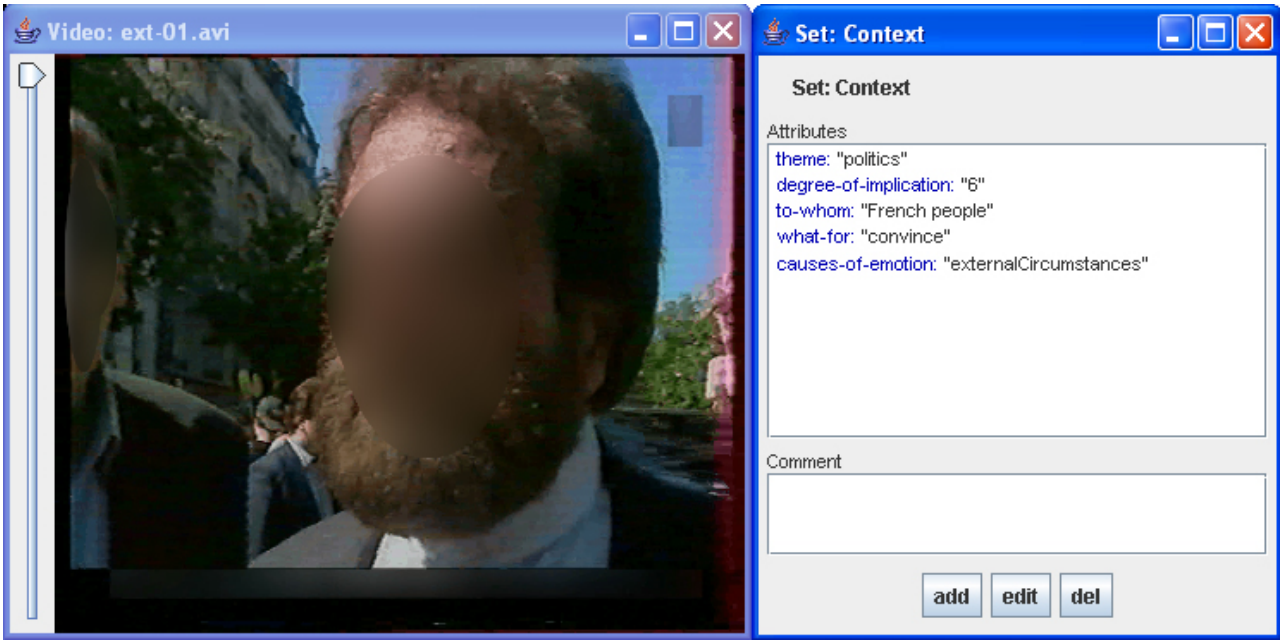

Figure 3: Context annotation with Anvil

4 Emotion annotation

In order to study the influence of the modalities on the perception of emotions, two annotators used the Anvil tool for annotating perceived emotions (Figure 4) in three conditions: 1) Audio only, 2) Video only, 3) Audio & Video.

4.1 Segmentation

The annotators were asked to create emotional segments where they felt it was appropriate (e.g. a period of time in the video where the emotional behavior was consistent). The two individually created set of segments were grouped into an agreed common set of emotional segments, as shown below (Table 3). The criteria for obtaining a unique segmentation were to use the union of the segments of the "audio only" condition on the one hand and the intersection of the segments of the "Video only" condition on the other hand. For the "Audio & Video" condition, the segmentation was done by Coder1, and Coder2 used the same segmentation.

Table 3: Number of segments of the whole corpus for each condition and coder, with agreed final segmentation

Condition	Number of segments of Coder1	Number of segments of Coder2	Final segmentation
Audio only	117	291	181
Video only	312	416	295
Audio & Video	234	343	281

We obtained more segments when the video was present. Indeed, audio emotional expressions (phrase or semantic/syntactic segments) are generally longer than visual ones, in a given video clip. Visual behaviors change more quickly (e.g. blinking). Visual modalities are also more numerous than audio ones. The use of the "Audio only" segmentation for the "Audio & Video" condition is not straightforward.

The second visible fact is that Coder2 annotated many more segments than coder1, for all the 3 conditions. This shows the difference of level of annotation of the two coders, Coder1 produced a small number of long segments, while Coder2 used a more precise segmentation with more segments of smaller duration.

4.2 Labeling

For labeling the emotion perceived in each segment, the two expert annotators selected a single emotional label of their choice (free text). The next table (Table 4) shows the number of labels annotated by each coder for the "Audio only" and "Video only" conditions.

Table 4: Number of labels of the whole corpus for each condition and coder

Condition	Number of different labels produced by Coder1	Number of different labels produced by Coder2
Audio only	76	38
Video only	86	70

Coder2 produced half the labels produced by Coder1 for the audio condition. All the 176 labels were classified into a set of 14 broad categories used for the annotation of the "audio and video" condition: anger, despair, disgust, doubt, exaltation, fear, irritation, joy, neutral, pain, sadness, serenity, surprise and worry. The coarse-grained level regrouping these 14 classes is composed of the 6 well-known Ekman classes plus the "neutral" class.

The annotators also used two classical appraisal dimensions (Cowie, Douglas-Cowie, Savvidou, McMahon, Sawey et al., 2000): intensity and valence (negative/positive emotion).

We included in our coding scheme both verbal categories and abstract dimensions in order to study their possible redundancy and complementarities. For example, redundancy between verbal categories (e.g. anger) and valence dimension (e.g. negative emotion) might be helpful for validating the annotations.

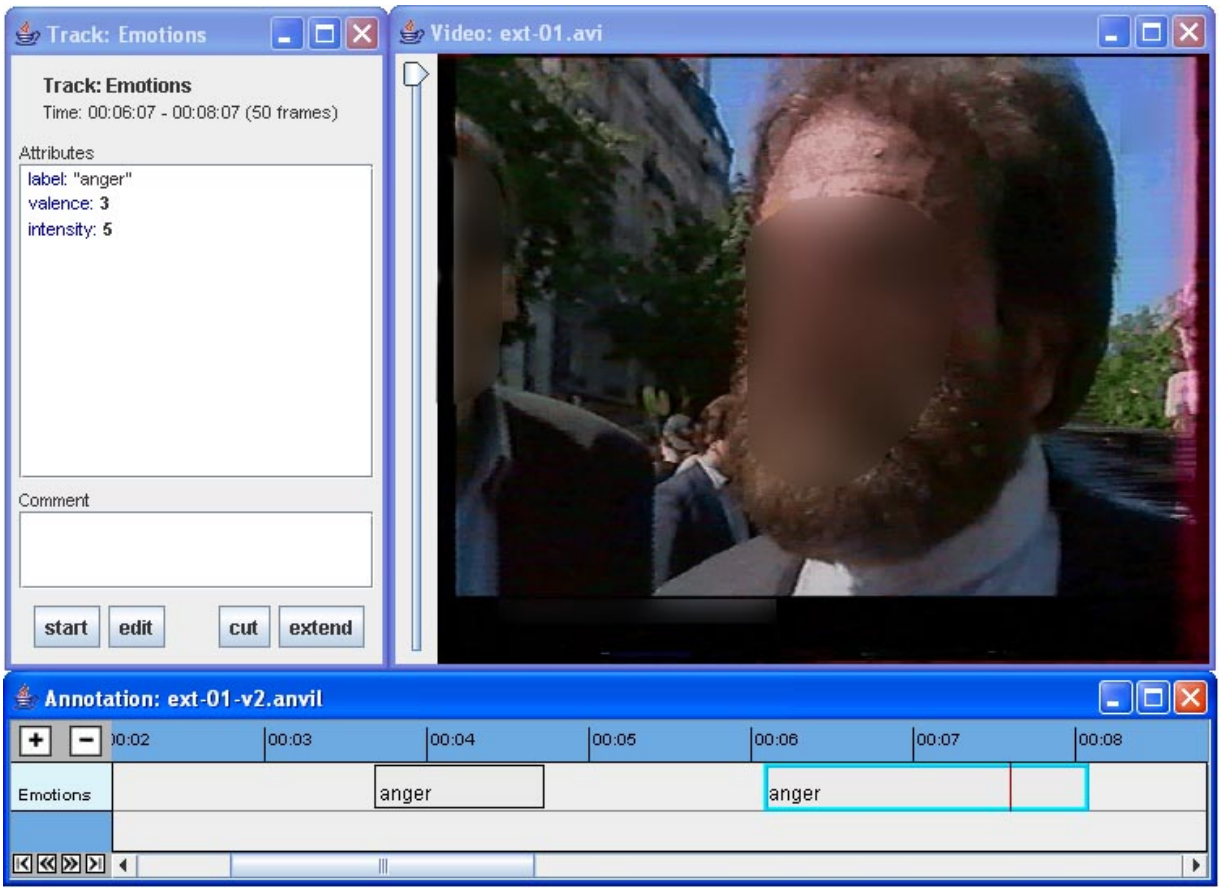

Figure 4: Emotion annotation with Anvil

5 Results of the annotation phase and discussion

Concerning the context, the free text coding of the "To whom" resulted in the following list: public, society, one persona, one group of persona, organization, and kin. The coders agreed mostly on the corpus, the "public" was often the value associated to this attribute; because the interviewed persons talk mostly to the camera, the emotional behavior is mostly directed to the people who will watch the interview. The attribute "Causes of emotion" could take one of the two values: external circumstances, or self behavior, and was mostly annotated as external circumstances.

Kappa inter-coder agreement measures for categorical variables (Carletta, 1996) have been computed on these manual annotations of emotions. Cronbach's alpha measures for continuous variables have been done on the valence and intensity of the emotions. The following table (Table 5) shows the results for each condition.

Table 5: Inter-coder agreement measures computed for each condition, on emotions, intensity and valence.

Condition	Kappa on the categories of Emotion	Cronbach's alpha on Intensity	Cronbach's alpha on Valence
Audio only	0.540	0.870	0.911
Video only	0.430	0.510	0.710
Audio & Video	0.370	0.254	0.574

Concerning the categories of emotion, the Kappa values were the lowest for "Audio & Video", then for "Video only" and higher for "Audio only". After classifying the intensity and valence values (going from 1 to 7), in 3 broad classes, for intensity: (1, 2, and 3 = low); (4 = middle); (5, 6, 7 = high), and for valence: (1, 2, and 3 = negative); (4

= indeterminable); (5, 6, 7 = positive). Results reveal a problem of balance between the perceptions of high and low intensities, in the case of “Audio & Video”; both coders annotated numerous “high” values for intensity (agreement on 144 “high” intensity annotations, but no annotation agreement for the “low” value for intensity). Results also reveal an equally balance between negative and positive emotional behaviors.

The next table (Table 6) shows, for each label or group of labels, the quantitative similarities, between both coders, on emotion annotations, for the 3 conditions.

Table 6: Quantitative similarities between both coders, on emotion annotations, for the 3 conditions (in percentage of annotation agreement)

Emotion\Condition	Percentage of annotation agreement for “Audio only”	Percentage of annotation agreement for “Video only”	Percentage of annotation agreement for “Audio & Video”
Anger	17%	20%	19%
Despair	6%	7%	5%
Doubt	4%	5%	0%
Disgust	1%	1%	0%
Exaltation	6%	5%	8%
Fear	0%	0%	1%
Irritation	8%	5%	6%
Joy	40%	35%	36%
Neutral	8%	1%	12%
Pain	2%	0%	7%
Sadness	7%	10%	3%
Serenity	0%	0%	3%
Surprise	0%	2%	0%
Worry	1%	9%	0%

The results show some tendencies: a high level of agreement for all conditions, concerning the labels "Anger", "Irritation”, "Exaltation", and "Joy". The coders also agreed, for "Surprise" and "Worry", mainly for the video condition. They agreed on the "Pain" attribute for the "Audio" and "Audio&video" conditions but not for the "Video" condition, which could be explained by the fact that acoustic cues show well this emotion (tears, cry). They agreed on the "Doubt" attribute only for the "Audio" or "Video" condition. The "Serenity" label, which represents a very subtle emotion, had agreement only in the "Audio & video" condition. The coders never agreed for "Neutral" for the "Video only" condition but also had a low level of agreement for the other conditions which shows the difficulty to define this behavioral state. The "Neutral" state is mostly used for annotating behaviors of very low intensity.

Valence values selected by the two annotators (Figure 5) were highly related for the three classes of labels : positive, negative and neutral.

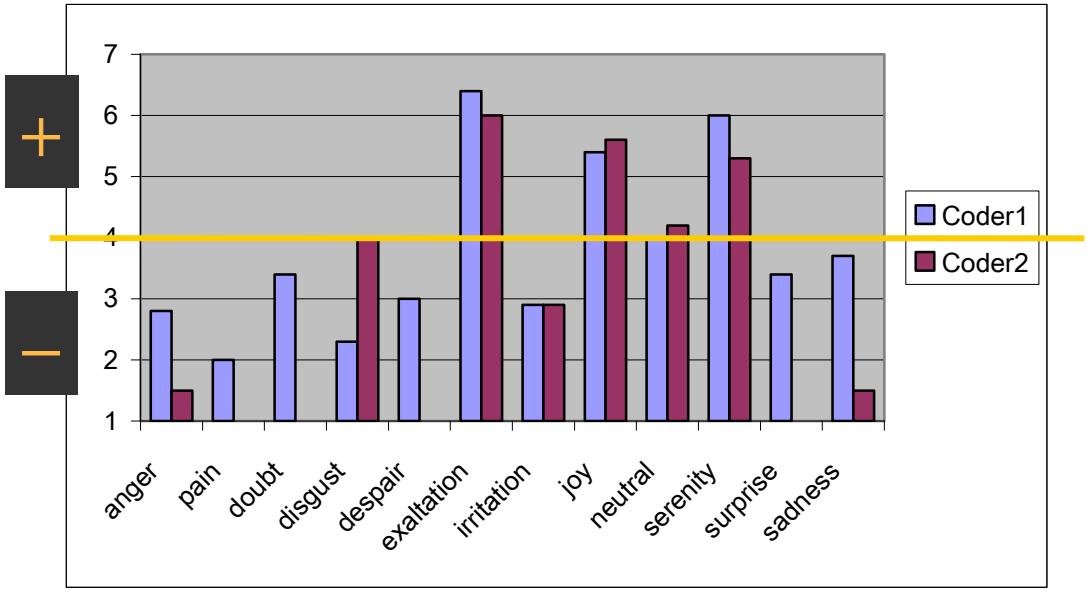

Figure 5: Valence and emotion relations in the case of "Audio and video" condition

The next figure (Figure 6) shows the label distribution of each coder. The most frequently used verbal labels by both coders revealed to be: "joy", "neutral", "sadness", "irritation", and "anger".

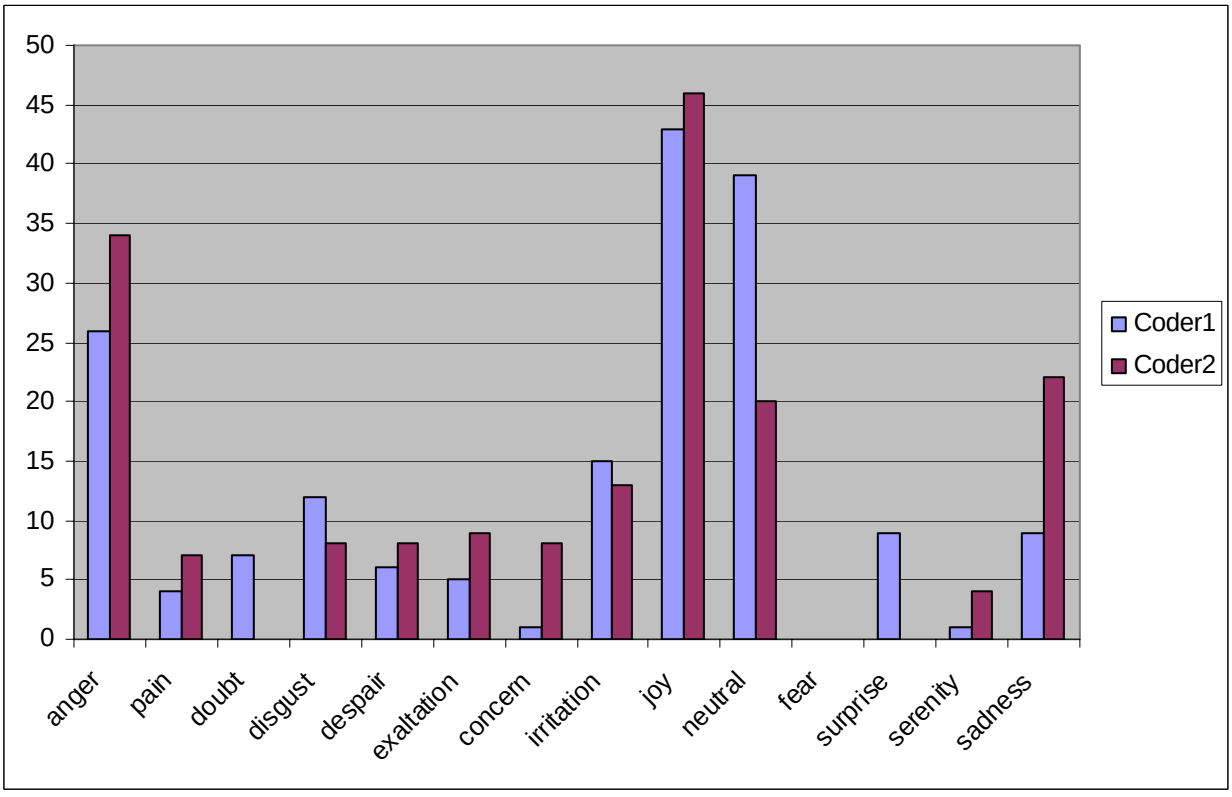

Figure 6: Label distribution of Coder1 and Coder2 in the case of "Audio and video", in percentage of annotation

Examples of inter-coder disagreement are listed below (Table 7), showing the high subjectivity of emotion perception. Disagreements can occur in the same valence class (Clip 3), and between positive/negative classes (Clip 4).

Table 7: Examples of inter-coder disagreements

Extract	Condition	Label annotated by Coder1	Label annotated by Coder2
Clip 3	Audio & Video (segment 1)	Anger	Sadness
	Audio & Video (segment 3)	Anger	Despair
Clip 4	Audio only	Sadness	Sadness
	Video only	Sadness	Neutral
	Audio & Video	Joy	Sadness

Ambiguities appear due to non-basic emotional patterns. The inter-coder agreements reveal to be lower than those obtained with acted emotions. We explain these low inter-coder agreements by the fact that real life emotions contain ambiguities and conflicting cues between emotions perceived in audio signal and in multimodal behaviors, but also within the visual channels. People in video often mask their emotions, try to control them, for example by smiling while talking with a trembling voice. The possibility to select only one label thus reveals to be not enough to take account of the many complex emotions and combinations that can occur in natural settings. Few disagreements were due to negative/positive confusion as shown below (Table 8).

Table 8: Negative / positive confusion percentages for each condition.

Condition	Negative / positive confusion percentages
Audio only	3%
Video only	7%
Audio & Video	11%

Indeed, the most frequent disagreements concerned fear/anger/sadness/disgust broad classes at different levels of intensity. In several clips, one might perceive sadness and anger at the same time, because of conflicting multimodal cues in speech and facial expression, or because of a transition between these two emotions. Furthermore, for such natural and complex emotional behaviors, two types of emotions may be perceived in the same time, the speaker emotion based on his/her internal emotional state, and the effect that s/he would be likely to have on the listener.

These observations of cases leading to inter-coder disagreements led us to propose the following typology of non basic emotional patterns:

- *Low-intensity emotion*: coder hesitates between "neutral" and the given label.
- *Blended emotions*: two emotions are mixed, and occur at the same time, often for emotions of the same valence, for instance sadness and anger.
- *Masked acted emotion*: the videotaped person is masking her real emotion, like a joy mask (by smiling) with a real disappointment behind.
- *Sequence of emotions*: two emotions, occurring one after the other, on a single emotional segment.
- *Cause-effect conflict*: positive/negative conflict for example (e.g. to cry for joy).
- *Emotion ambiguity*: it is difficult or impossible to decide between two emotions.

6 Future directions

This study provides both concrete suggestions (the set of labels for annotating emotional behaviors and the typology of non basic emotional patterns) but also long-term directions for the design of multimodal affective interfaces. The results described above show that the possibility to select only one label is not enough. We thus modified our coding scheme by allowing the annotation of each emotional segment with two labels, and by extending the emotional labels list from 14 to 18 labels, the new list is the following: anger, despair, disappointment, disgust, doubt, embarrassment, exaltation, fear, irritation, joy, neutral, pleased, pride, sadness, serenity, shame, surprise, worry. Three other expert annotators will annotate the same set of videos in the audio-video condition with this new coding scheme. The corpus will be extended, with more videos, and new annotations will be done with the new emotion context coding scheme. We have also defined a scheme for coding multimodal behaviors. The results of the first

annotation phase are described in (Abrilian, Martin & Devillers, 2005). This multimodal coding scheme was validated and corrected after this first annotation phase. It will enable the modeling of the relations between emotional labels and multimodal annotations. Such a model will be useful for the specification of Human-Computer Interfaces involving affective Embodied Conversational Agents.

Acknowledgments

The work described in this paper has been partly funded by the FP6 IST HUMAINE Network of Excellence (http://emotion-research.net). The authors thank Marianne Najm and Fabien Bajeot for their annotations.

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. *Proceedings of HCI International 2005*, 22 - 27 July.

Batliner, A., Fisher, K., Huber, R., Spilker, J., & Noth, E. (2000). Desperately seeking emotions or: Actors, wizards, and human beings. *Proceedings of SpeechEmotion-2000*, pp. 195-200.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2), 249–254.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). Embodied Conversational Agents: MIT Press.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time. In *ISCA Workshop on Speech & Emotion*, (pp. 19-24). Northern Ireland.

Craggs, R., & Wood, M.M. (2004). A categorical annotation scheme for emotion in the linguistic content. *Proceedings of Affective Dialogue Systems (ADS'2004)*.

Devillers, L., & Vasilescu, I. (2004). Reliability of Lexical and Prosodic cues in two real-life spoken dialog corpora. *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC'2004)*.

Douglas-Cowie, E., Cowie, R., & Schröder, M. (2000). A New Emotion Database: Considerations, Sources and Scope. In E. Douglas-Cowie, Cowie, R. and Schröder, M. (Ed.), *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, (pp. 39-44).

Douglas-Cowie, E., Campbell N., Cowie R., & Roach P. (2003). Emotional speech: towards a new generation of databases. *Speech communication*, 40.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M.J. Power (Eds.), *Handbook of Cognition & Emotion*, (pp. 301–320). New York: John Wiley.

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of Eurospeech'2001*.

Martin, J.-C., den Os, E., Kuhnlein, P., Boves, L., Paggio, P., & Catizone, R. (2004). Workshop "Multimodal Corpora: Models Of Human Behaviour For The Specification And Evaluation Of Multimodal Input And Output Interfaces". *Proceedings of In Association with the 4th International Conference On Language Resources And Evaluation LREC2004 http://www.lrec-conf.org/lrec2004/index.php*, 25th may.

Paiva, A. (2000). Affective Interactions, Towards a New Generation of Computer Interfaces. Berlin: Springer-Verlag.

Plutchik, R. (1994). The psychology and Biology of Emotion. New York: Harper Collins College.

Scherer, K.R. (2000). Emotion. In M.H.W. Stroebe (Ed.), *Introduction to Social Psychology: A European perspective*, (pp. 151-191): Oxford: Blackwell.

Wegener Knudsen, M., Martin, J.-C., Dybkjær, L., Berman, S., Bernsen, N.O., Choukri, K., Heid, U., Kita, S., Mapelli, V., Pelachaud, C., Poggi, I., van Elswijk, G., & Wittenburg, P. (2002). *Survey of NIMM Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources. ISLE Natural Interactivity and Multimodality. Working Group Deliverable D8.1*.