Creation of a Corpus of Multimodal Spontaneous Expressions of Emotions in Human-Machine Interaction

Le Chenadec G.[1], Maffiolo V.[1], Chateau N.[1] and Colletta J.M.[2]

[1]France Telecom R&D, Technologies Division, Technopole Anticipa, 2 av. Pierre Marzin, 22307 Lannion, France

[2] Lidilem, Université Stendhal, BP25, 38040 Grenoble Cedex 9

E-mail: {gilles.lechenadec, valerie.maffiolo, noel.chateau}@francetelecom.com, jean-marc.colletta@u-grenoble3.fr

Abstract

This paper presents an experience in laboratory dealing with the constitution of a corpus of multimodal spontaneous expressions of emotions. The originality of this corpus resides in its characteristics (interactions between a virtual actor and humans learning a theater text), in its content (multimodal spontaneous expressions of emotions) and in its two sources of characterization (by the participant and by one of his/her close relation). The corpus collection is part of a study on the fusion of multimodal information (verbal, facial, gestural, postural, and physiological) to improve the detection and characterization of expressions of emotions in human-machine interaction (HMI).

1. Introduction

Everyday-life automated systems have often difficulties to optimally handle interactions with humans. Due either to the quality of the media or the variability of human behaviour and characteristics, pseudo-intelligent systems may fail to recognize users' request. A way to regulate the interaction or to improve speech recognition unit is to detect user's affective states. On another side, technological progress provides more and more functionalities allowing communicating with the machine (audio, video, haptic…). Computer systems which are able to recognize human emotions or affective states from speech, postures, gestures or other modalities may enhance HMI, assuming a strategy to adequately answer has been determined. Our work focuses precisely in this context: the detection of human multimodal expressions of emotions in the human-machine interaction.

The basis of this work is the corpus collection. In the literature (Douglas-Cowie *et al.*, 2004), several corpuses of multimodal expressions of emotions or affective states exist but either emotions are acted (non-spontaneous), or recordings are limited to one or two-modes only, or the corpus is not obtained in a human-machine interaction, or no emotional labels are available. The corpus collection presented in this paper is the first step of a study on the fusion of multimodal (verbal, facial, gestural, postural, and physiological) and spontaneous expressions of emotions.

In section 2, some key points for a corpus collection with emotional content are developed. In the two next sections, methodologies intended to elicit expressions of emotions and to obtain their characterization, are described. Section 5 gives a global overview of our experimental setup. First analyses are set out and discussed in section 6 and conclusions are presented in section 7.

2. Some Key Points

As stand out in (Douglas-Cowie, 2004), key points for collecting databases with an emotional content, are the identification of target emotions, the choice of a method for expression elicitation, recording modalities and the database labelling. These key points are discussed thereafter.

In databases with expressions of emotions, the set of target emotions has been first limited to "primaries" or "big six" emotions (Ekman and Friesen, 76). Recently, researchers focused on more complex emotions which describe more precisely the range of everyday-life emotions. This includes emotions-related states (Cowie and Cornelius, 2003) or mental states (Baron-Cohen *et al.*, 2004).

A wide range of methods is used to elicit expressions of emotions from which three types emerge. Asking for acted expressions is the first one (Banse and Scherer, 1996; Polzin and Waibel, 2000; Baron-Cohen *et al.*, 2004). A study by Batliner *et al.* (2003) shows that the results obtained via this methodology cannot be transposable to everyday-life. For an objective of detection, there is a gap between acted and spontaneous expressions of emotions. The collection of naturalistic data constitutes the second type of elicitation method. In this one, recording expressions of emotions are those which may be occurred in the everyday-life. For instance, EmoTV1 database (Abrilian *et al.*, 2005) contains audio-video recordings of people in TV interviews. The third type of elicitation method allows recordings of induced data in laboratory conditions which are not acted expressions of emotions. Advantages of this last type stand in the quality of recorded signals and the control of stimuli (Aubergé *et al.*, 2003). In particular, this method often includes the Wizard-of-Oz paradigm allowing simulating HMI whereas researchers control the machine.

Researches on the constitution of databases containing multimodal expressions of emotions are very recent. Moreover, they often focuses on the recording of two modalities of expression, facial and speech ones being the most studied e.g. the Belfast Naturalistic Database (Douglas-Cowie et al., 2000) and the SALAS database (http://www.image.ntua.gr/ermis/). Very few deals with more than two modalities: ORESTEIA database (McMahon et al., 2003) contains speech, facial and physiological measurements and SMARTKOM (Schiel *et al.*, 2002) records speech, facial expression and gestures.

The last key point concerns the labelling of recorded expressions of emotions that can be split in "encoding emotion" and "encoding the signs of emotions". Encoding

emotion consists in a choice between a discrete categorical approach (Abrilian *et al.*, 2005) and a continuous dimensional approach (Douglas-Cowie *et al.*, 2000). Note a recent collaboration (Douglas-Cowie *et al.*, 2005) that gathers experience mixing categorical and dimensional approaches, showing robustness of a complementary methodology to determine roles of audio-visual modalities. Encoding signs of emotions provides relevant high-level data to compensate either the fact that automatic extraction of useful descriptors of each modality is not yet optimal. For facial signs, Ekman and Friesen (1978) developed a standard method of labelling called the Facial Action Coding System (FACS). For other modalities, no standard exists. Speech expressions are those that have been the more widely studied.

The next two sections discuss these key points and set out the methodologies developed for the test.

3. Elicitation Methodology

In the aim of collecting a multimodal spontaneous corpus, the first objective was to elicit a wide range of emotions characterized by either positive or negative valence. Moreover, elicited expressions had to be spontaneous (opposite to acted expressions). In the optic to develop affective computer systems which detect and characterize expressions of emotions, our database has to reflect the multimodal character of human behaviour.

A laboratory test platform has been developed allowing us to handle an interaction between a human and a virtual character based on the Wizard-of-Oz methodology. In this interaction, participants think they interact with an autonomous system whereas it is controlled by researchers. As mentioned previously, this methodology combined with the developed platform has permitted a precise control of recording quality and an efficient course of elicitation methodology.

The success of the collection depends on the relevance of the chosen methodology (application and scenario of interaction) to elicit spontaneous expressions of emotions. The main difficulty consists in imagining an application and a scenario of interaction in which participants imply themselves in the dialogue. In addition, the choice of the application and the scenario has to take into account the variability of participants' reactions, emotional or intentional. Inspired by a previous experiment (Chateau *et al.*, 2005), the application used is based on the interaction between a virtual actor playing opposite a real human actor (the participant). The scenario consists for the participant in learning three scenes of *Don Quixote de la Mancha* written by *M. de Cervantes* in 1605. The Facelab application (Breton *et al.*, 2001; Courty *et al.*, 2003) is used to control the virtual actor's mimics, head movements and speech. Cues of the virtual actor are controlled by a researcher in real time, simulating an autonomous system. The Experimental setup is described in section 5.

Beyond acted expressions of emotions required by the role, the developed application simulates bugs of the system during interaction to elicit spontaneous expressions of emotions. Different bugs have been designed which involve the synthesis unit (uncoordinated movements or stammering of the virtual actor), the speech recognition unit (the system asks several times the real actor to repeat his/her cue) or the global system unit (the system displays "Lost data", a researcher asks the real actor to play again the scene from the beginning). Note that for the second type of bugs, participants cannot know if repeating the line is due to their play or to a system's bug. Other types of bugs are clearly related to a system's failure in the participant's point of view.

The first phase of the test is dedicated to the instructions reading and the calibration of sensors (video cameras, microphones, and physiological sensors like blood volume pulse and skin conductance). It is told to participants that they have to test a new application aiming to rehearse and to get *Don Quixote*'s cues learned. Participants think that the system operates autonomously and that sensors allow the system to recognize speech (lavalier microphone) and to detect position and gesture movement of participants (cameras and finger sensors).
The second phase concerns the interaction between the participant and the virtual actor. This interaction has two steps: in the first one, the real actor plays opposite the virtual actor. In the second step, it has been added in the instruction that the system is capable of judging the real actor's play. In this step, system bugs are launched by the researcher.

4. Labelling Methodology

In this collection, we strive to enhance labelling quality both in the characterization of recorded emotions and the identification of emotional signs.

After the phase of interaction between the two actors, a phase of labelling is conducted. In this phase, the recording of the interaction (from the real actor's viewpoint, in which only the virtual actor is visible) is played back to the participant in order to make him/her live back his/her interaction with the virtual actor and to make him/her comment precisely and freely emotions he/she felt. The recording of the interaction is displayed twice. In the first display, the participant has to characterize emotions in few words and starting time of what he/she felt during the interaction. In the second, he/she determines the ending time of his/her feeling.

Subsequent interviews are conducted with a close relation of the real actor (husband, wife, friend, etc.). This original point of methodology allows to access to precise labelling (both on emotions or signs of emotions) which will be useful to enhance detection of expressions of emotions.
During the interview with the close relation, the recording of the interaction (from an external viewpoint, in which both actors are visible) is played back. The close relation is asked to characterize emotions expressed by the participant, to describe signs of emotion on different modalities (speech, facial mimics, gestures, and postures) and to determine time boundary of each emotion.

5. Experimental Setup

According to the elicitation and labelling methodologies, the experimental setup is distributed in three rooms: the first one is dedicated to the interaction with the application,

the second one is dedicated to the system's control and the recording of the sensors, and the last one is used for labelling interviews.
During the interaction, participants stand alone in the room rehearsing with the virtual actor (Figure 1). A multimedia set (TV + speakers) displays the animated virtual actor and real actor's cues (prompter) and handles sound restitution. Three commercial DV cameras record video signals of the interaction. The two first ones are placed above the TV set (Figure 1) and record facial mimics and gestures/postures of the real actor (Figure 2). The last DV camera records the TV screen for the labelling interview with the participant.

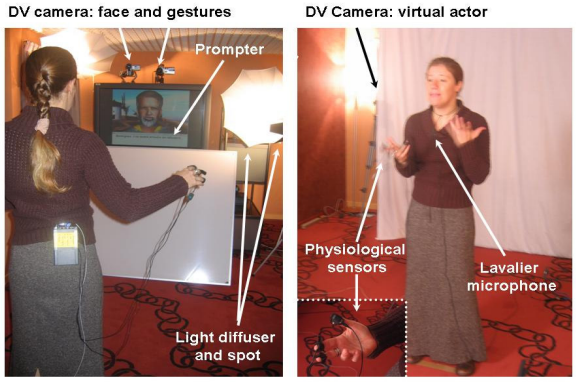

Figure 1. A participant (left is rear side and right is front side) in interaction with the virtual actor.

Audio signal is recorded with a lavalier microphone clipped on the participant (Figure 1). Three skin sensors are set on fingers (Figure 1) and record skin conductance, blood volume pressure and heart rate.

Figure 2. Examples of video signals from DV cameras recording face (left), gestures and postures (middle) of the real actor and the virtual actor (right).

The system's control room gathers materials for the control of human-machine interaction and signal digitization of each sensor. A computer with two video outputs handles the Wizard-of-Oz interaction. The first output displays the prompter and *Facelab* commands. The second output is displayed on two screens. The first screen allows a visual control for researchers and the second one is seen by the participant. Signals recorded from the lavalier microphone are digitized by a DAT. Physiological signals are directly digitized by a dedicated computer.
The technical setup of the labelling room is composed of a TV set and a video-tape to read and display the interaction. An additional camera records the interview.

6. First Analysis and Discussion

Nine females and nine males, aging from 25 to 50, and eighteen close relations took part in the experiment. Tests with participants (pre-test and test) lasted two months. Labelling interviews with close relation have been conducted later during three months. Interaction durations are between thirty and forty-five minutes. This variability is essentially due to either technical problem during interaction or difficulties to repeat lines (stammering, uncontrollable laugh). For each participant, recorded data gather forty-minutes (mean time) of mimics and body video, audio signal and physiological data. Interview recordings lasted 1h15m for each participant and 2h for each close relation.

At the end of the experiment and before any global analysis, a first data observation reveals that the corpus contains numerous fine variations of spontaneous facial mimics, postural and physiological expressions, but a lack of occurrences of verbal and gestural expressions appears. This lack of spontaneous verbal or gestural expressions is likely due to the fact that participants knew they took part in an experiment. First, the interaction with the virtual actor occurred in a laboratory, and participants knew they were filmed. Moreover they were recruited for a serious task and they did not bought the application tested (probably they would have expressed more emotions if they would have been at home having spent money with an application working with such bugs). Second, the virtual actor does not use gestures to express itself and participants acts in the same way. Concerning the lack of speech expressions, we may think that the interaction based on speech recognition does not lead participants to verbalize spontaneously. Lastly, some participants played their part and did not get right outside despite the system's bugs (supposed to elicit expressions of emotions).
Nevertheless, some participants' behaviours show a complete multimodal quality content.

Labelling interviews (both participants and close relations) provide precise characterization of expressions either about the emotions or the signs of emotions. More than emotions labelling, first analyses of these interviews reveal emotion-related states (e.g. helplessness) and cognitive states (e.g. wondering). Main occurrences of emotions are relative to anxiety, stress, irritation, amusement, incomprehension, boredom, relief: showing the diversity of emotions of participants during the interaction.

To complete labelling of expressions of emotions recorded in the interaction phase, interviews will be carried out later with a group of third party observers. The matching between the three sources of emotional labelling is intended to bring fruitful information to stand out our strategy to develop tools for detecting expressions of emotions. A precise labelling of expressions of emotions is necessary in order to enhance modelling of human behaviour. In this corpus, three sources of characterization of signs will be available. The participant gives his/her feelings, close relation and third party observers give feelings they supposed the participant felt respectively with and without a significant degree of familiarity. Note that no sign of expressions of emotion is cited and characterized by participants. First, we cannot ask participants to reactivate their emotions and also analyse their expressions. Moreover, they are not regular in the observation of them and close relation will provide more precise information.

We can reasonably suppose that close relations and third-party observers can detect easily cognitive states. Inference of emotion-related states may necessitate either acquaintance or standard expressions coming from the participants. The study is intended to answer these questions and bring significant information about the nature of recorded expressions.
The two sources of information (close relation and third-party observers) furnish two different viewpoints on participants' behaviours: one specific and restricted (intimate); one more general and representative of an anonymous judgement. A complementary study will investigate the relation between the quality of modelling and these two sources of information.

7. Conclusion

This paper describes the creation of a corpus of multimodal spontaneous expressions of emotions in Human-Machine Interaction. Two methodologies have been elaborated in order to elicit and label expressions of emotions. Compared with other elicitation methodologies, the scenario used is based on an original interaction between a real human actor and a virtual animated actor. Labelling from close relations' point of view provides a precise characterization of signs.
First analyses show that the content of the corpus is rich but not complete for all the aimed modalities. Subtile expressions of emotion-related and cognitive states have been recorded. The diversity of the corpus content is a great starting point for the study of the detection of the detection and the characterization of multimodal expressions of emotions.

8. References

Abrilian S., Devillers L. & Martin J.C. (2005). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. In proceedings of HCI International, Las Vegas.

Aubergé V., Audibert N. & Rilliard A. (2003). Why and how to control the authentic emotional speech corpora. In proceedings of EuroSpeech, Geneva.

Banse, R. & Scherer, K. (1996). Acoustic profiles in vocal emotion expression. Journal of personality and social psychology 70 (3), 614-636.

Baron-Cohen S., Golan O., Wheelwright S. & Hill J.J. (2004). Mind Reading: The Interactive Guide to Emotions. London: Jessica Kingsley Publishers.

Batliner A., Fischer K., Huber R., Spilker J. & Nöth E. (2003). How to find trouble in communication. Speech Communication 40, 117-143.

Breton G., Pelé D. & Bouville C. (2001). FaceEngine : a 3D facial animation engine for real time applications. In proceedings of ACM WEB3D, Paderborn, Germany.

Chateau N., Maffiolo V., Pican N. & Mersiol M. (2005). The Effect of Embodied Conversational Agents' Speech Quality on Users' Attention and Emotion. In proceedings of ACII (pp. 652—659), Beijing, China.

Courty N., Breton G. & Pelé D. (2003). Embodied in a look: bridging the gap between humans and avatars. In proceedings of IVA 03 "Intelligent Virtual Agents", Irsee, Germany.

Cowie R. & Cornelius R. (2003). Describing the emotional states that are expressed in speech. Speech Communication 40, 5-32.

Douglas-Cowie E., Cowie R. & Schröder M. (2000). A new emotion database: considerations, sources and scope. SCA Workshop on Speech and Emotion, Northern Ireland.

Douglas-Cowie E. & WP5 members (2004). HUMAINE deliverable D5c: Preliminary plans for exemplars: databases, http:// emotion-research.net/deliverables.

Douglas-Cowie, Devillers L., Martin J-C., Cowie R., Savvidou S., Abrilian S. & Cox C. (2005). Multimodal databases of everyday emotion: content and labeling. In proceedings of Interspeech'05.

Ekman P. & Friesen W.V. (1976). Pictures of Facial Affect. Consulting Psychologists Press.

Ekman, P. & Friesen, W. (1978). The Facial Action Coding System. Consulting Psychologists' Press, San Francisco, CA.

McMahon E., Cowie R., Kasderidis S., Taylor J. & Kollias, S. (2003). What chance that a DC could recognise hazardous mental states from sensor outputs? In proceedings of Tales of the Disappearing Computer, Santorini.

Polzin, T. S. & Waibel, A. (2000). Emotion-sensitive human-computer interfaces. In proceedings of ISCA ITRW on Speech and Emotion, Newcastle, 5-7 September 2000, Belfast, pp. 201- 206.

Schiel F., Steininger S. & Türk U. (2002). The SmartKom Multimodal Corpus at BAS. In proceedings of LREC 02, Las Palmas, Gran Canaria, Spain, pp. 200-206.