Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions

Fabien Ringeval[1], Andreas Sonderegger[2], Juergen Sauer[2] and Denis Lalanne[1]

Abstract— **We present in this paper a new multimodal corpus of spontaneous collaborative and affective interactions in French: RECOLA, which is being made available to the research community. Participants were recorded in dyads during a video conference while completing a task requiring collaboration. Different multimodal data, i.e., audio, video, ECG and EDA, were recorded continuously and synchronously. In total, 46 participants took part in the test, for which the first 5 minutes of interaction were kept to ease annotation. In addition to these recordings, 6 annotators measured emotion continuously on two dimensions: arousal and valence, as well as social behavior labels on five dimensions. The corpus allowed us to take self-report measures of users during task completion. Methodologies and issues related to affective corpus construction are briefly reviewed in this paper. We further detail how the corpus was constructed, i.e., participants, procedure and task, the multimodal recording setup, the annotation of data and some analysis of the quality of these annotations.**

I. INTRODUCTION

The study of the complex phenomena portrayed by humans during social interactions, requires rich sets of labelled data of repeatable experiments with situations occurring in daily-life [1]. Such datasets enable researchers to have a better understanding of the relationships that may exist between behavioral cues, e.g., facial, vocal and gestural expressions, and their communicative functions during social interaction, e.g., engagement, dominance, emotion. This identification is usually achieved by the development of systems that exploit features extracted from sensors such as audio, video, motion or physiological signals, to detect and recognize automatically some components of the social signal [2]. When such systems perform well enough, they can be used in computer-aided analysis of human-human interactions, like socio-emotional analysis of married couples interactions [3], or individuals on the autism spectrum as they may have difficulties understanding them [4].

Although the last decade has seen a growing interest for collecting data of social interactions [5], [6], there is still a lack of spontaneous, socially rich and multimodal corpora. Spontaneous interactions are considered ideal for validating real-life affective analysis systems. Although such authentic behaviors are difficult to collect because they are relatively rare, short lived, and filled with subtle context-based changes [7]. Multimodal recordings that include physiological data such as, electrocardiogram (ECG) or electrodermal activity (EDA), may help to recognize the natural behaviors [8]. The strategy can, for example, be adapted with the availability of the modalities along the interaction course, e.g., no speech production and no visible face. Additionally, measurement techniques of both ECG and EDA signals from video data [9], which allow to remove the intrusiveness aspect of the typical biosignals recording devices, can be optimized using high-quality recordings of these data as ground-truth.

Several challenges must be faced when creating a corpus of social interactions [6], [7]. One of these challenges concerns the choice of the tasks that participants have to perform. The chosen interaction scenario should be easy to reproduce and taken out from those occurring in daily-life, to ensure having fully natural and socially rich interactions between participants. Concerning the interaction, it is known that both mood and emotion of a participant are influenced by those from others participants when they are interacting [10]. Influencing the participants' mood before the experiment, by imposing for example a specific context of interaction, may thus be useful to ensure having a good variety of behaviors during the interactions [11].

Another challenge that has to be faced when constituting a corpus concerns the annotation of the social behaviors from the multimodal recordings. Several models exist to explain the behavior a human may depict during social interactions, and how these behaviors may vary from person to person due to idiosyncrasies. The choice of a specific model will therefore influence the type of information used for judging the social signal. Consequently, relevant behavioral cues that might be identified by a recognition system depend on the chosen annotation scheme. Moreover, the way the information is being quantified (e.g., discrete or continuous time), and by who (e.g., annotators having a same or a different mother tongue than participants), can influence the resulting annotation, beside the fact that humans present biases and inconsistencies in their perceptual judgments [12].

In this paper, we introduce a new multimodal corpus of spontaneous interactions in French called RECOLA, for REmote COLlaborative and Affective interactions. The motivation for building this corpus is driven by the EmotiBoard project, which consists in the development of real-time automatic emotion recognizers to augment remote collaboration with emotional feedback, and measure the impact of such feedback on teamwork quality and efficiency. Since we could not find a corpus to train our models in the same remote settings and with multiple modalities, we decided to create one. The corpus was designed in a collaborative group constituted of psychologists and computational science researchers, and is being released in conjunction with this paper[1]. It conforms with many desirable criteria suggested by emotion researchers [13] and thus may serve as a valuable common resource for future studies on automated analysis of socio-emotional behaviors, as well as for those related to emotion contagion and interpersonal synchrony [14]. The particularities of this database, according to other emotional corpora, are as follow:

(i) it is based on spontaneous interactions collected from a collaborative task that was performed remotely, and where mood of participants was manipulated and balanced in dyads,

This work was supported by the IM2 Swiss National Centers of Competence in Research, a research instrument managed by the Swiss National Science Foundation on behalf of the Federal Authorities

[1] F. Ringeval and D. Lalanne are with the Department of Informatics, Université de Fribourg-Universität Freiburg, 1700 FR Fribourg, Switzerland

[2] A. Sonderegger and J. Sauer are with the Department of Psychology, Université de Fribourg-Universität Freiburg, 1700 FR Fribourg, Switzerland

Email:{fabien.ringeval,denis.lalanne}@unifr.ch

[1] Available: http://diuf.unifr.ch/diva/recola

TABLE I

OVERVIEW OF PUBLICLY AVAILABLE DATABASES WITH TIME CONTINUOUS LABELING OF EMOTIONALLY COLORED CONTENT

Database	Nr. Part.	Duration	Inter. Scen.	Recordings	Video Bandwidth	Annotators
Vera am Mittag [1]	20	12:00	Spontaneous	Audiovisual	352x288, 25Hz	17
SAL [5]	4	04:11	Induced	Audiovisual	352x288, 25Hz	4
SEMAINE [11]	20	06:30	Induced	Audiovisual	580x780, 50Hz	2-8
RECOLA	46	03:50	Spontaneous	Audiovisual, ECG, EDA	1080x720, 25Hz	6

(ii) it contains multimodal data, with detailed annotation of emotional (continuous time and valued scale) and social (discrete time and valued scale) primitives evaluated from both internal and external views, i.e., from participants and annotators, respectively,

(iii) and all the annotators annotated all the sequences, which guarantees a consistency in the annotations.

In the remainder of this paper we introduce some related works on emotional corpora construction (Sec. 2), then present the corpus construction (Sec. 3), with the multimodal recording setup (Sec. 4) and the annotation procedure with a short statistical analysis of the data (Sec. 5) before concluding (Sec. 6).

II. RELATED WORKS

There has been a considerable amount of work during the last decade for creating reusable multimodal corpus consisting in affectively and socially enriched human behavior. The latest and most extensive overview of emotional corpora can be found in the HUMAINE database [5], whereas those devoted to others social behaviors are reviewed in [6].

Three main types of interaction scenarios have been used to record emotionally colored interactions: (i) posed behavior [15], which is produced by the subject upon request, e.g., actors, (ii) induced behavior [5], [11], which occurs in a controlled setting designed to elicit an affective reaction such as when watching movies and (iii) spontaneous behavior [1], which appears in real-life settings such as interviews or interactions between humans or between humans and machines. Interaction scenarios with a posed behavior are the easiest to design and present the advantage to have a control on the portrayed emotions. However, this approach was criticized for including (non-realistic) forced traits of emotion, which are claimed to be much more subtle when the emotion arises from a real-life context [16]. Existing differences between acted and non-acted emotional speech are not yet clearly identified and need further investigation. Scenarios based on the induced behavior also permit indirect control of the behaviors of participants, by imposing a specific context of interaction, e.g., four emotionally stereotyped conversational agents were used in [11]. However, this approach may not provide fully natural behaviors, because the interaction may be restricted to a specific context, e.g., users were not allowed to ask questions and agents had to use key phrases in certain scenarios in [11], wherein the spontaneous aspect of interaction may be thus limited or even absent [17]. Finally, the spontaneous behavior scenario guarantees natural emotionally colored interactions, since the displayed behaviors arise from a real-life context where the set of produced verbal and non-verbal cues is both free and unlimited. However, this spontaneous interaction scenario is the hardest to design as it includes several ethical issues, like people discussing about private things, or not knowing they are recorded. Moreover, the affective behaviors cannot be properly controlled, even if there do exist mood induction techniques that can be easily used in laboratory settings, but much less in a fully natural environment.

Concerning the annotation scheme, several models can be used to quantify affective behaviors: (i) a categorical model, where an item is chosen from a list of affective related words, e.g., the full-blown emotions defined by Ekman et al. [18], to judge the content of a fixed length stimuli, (ii) a dimensional model [19], where a value is chosen over a time-continuous emotional scale during the play of a stimuli, like in Feeltrace [20] and (iii) an appraisal-based model, where the emotion is quantified with various stimuli evaluation checks, such as the novelty, intrinsic pleasantness, goal-based significance, coping potential and compatibility with standards [21].

Each approach, categorical, dimensional or appraisal, has its advantages and disadvantages [7], [22]. In the categorical approach, mixed emotions can not adequately be transcribed into words, which are moreover restrictive and culturally dependent [23]. Whereas in dimensional approach, observers can express their perception of the affective stimuli on several time scales, such as arousal, i.e., how excited or apathetic the emotion is, and valence, i.e., how positive or negative the emotion is [19]. But the reduction of emotion space to two or three dimensions may be too extreme and resulting in loss of information [24]. Concerning the appraisal-based approach, how to use it for quantifying affective variability remains an open research question as it requires complex, multi-componential and sophisticated measurments of change [22].

All of the aforementioned issues, and in particular the issue of which psychological model of emotion is more suitable for which context, still remain under debate [21]. Even though both arousal and valence dimensions are not claimed to be sufficient for differentiating all possible emotions, we have chosen to use them for annotating the affective behaviors collected in the RECOLA corpus, since they have proven to be useful in several domains related to affective content analysis [25]. Further, there does not exist to our best knowledge any freely available corpus that includes recordings of spontaneous behaviors with both audiovisual and physiological data, cf. Table I. Even if more than one communication modality can be studied from a same information source, e.g., gestural and facial expression analysis from video data [22], we believe that fusion of multimodal data, i.e., audio, visual and physiological, can help facing the complexity of automatic analysis of spontaneous affective behaviors [26].

Another novelty of our multimodal corpus concerns the chosen scenario: remote collaborative teamwork. Previous research in the domain of work psychology has indicated that mood, emotion and team members empathy may influence team processes and the outcomes of teamwork, such as performance, cohesion and satisfaction [27]. This suggests that awareness of each team members emotional state is important for efficient and satisfactory teamwork. Since research in the domain of affective computing has made significant progress in automatically detecting emotional states of humans [7], by analyzing speech behavior, facial expressions or physiological data, we planned to develop a tool, the EmotiBoard, for providing automatic affective feedback in teamwork during video-conference.

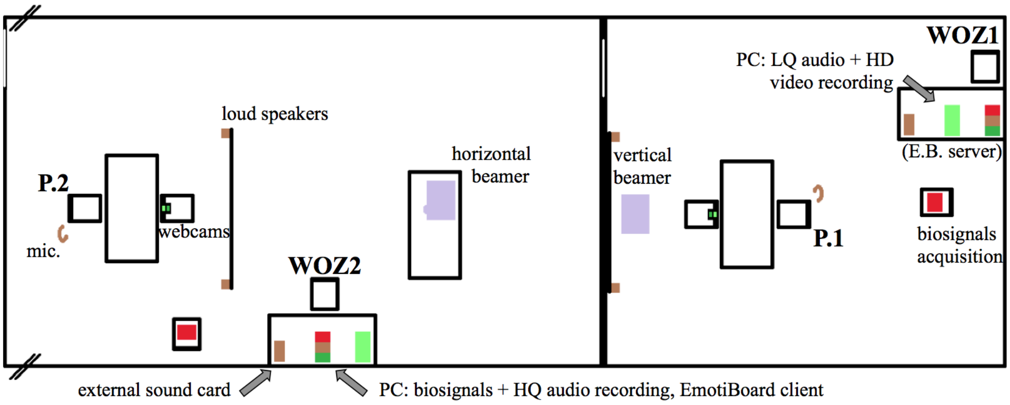

Fig. 1. Experiment setting used in the RECOLA corpus to collect multimodal data of spontaneous interactions in remote condition.

This is the reason why we recorded the RECOLA corpus in a remote collaborative setting, to have a ground-truth to build emotion recognizers. Further, we believe this scenario is relevant for a socio-emotional analysis prospective, since it conveys a spontaneous context of interaction. Following the work of [11], we decided to use mood induction techniques for collecting affective data in laboratory settings, since they ensure a variety of behaviors by participants, without introducing the ethical issues met in a fully natural environment, i.e., people wearing sensors 24/7. This mood induction procedure was intended to increase the difference in emotional valence between participants of a team, while slightly increasing their arousal.

III. CORPUS CONSTRUCTION

A. Participants and environment

46 participants (27 females, 19 males) from the department of psychology of the Université de Fribourg-Universität Freiburg, Switzerland, were recruited as 23 dyadic teams work. The master degrees issued in this university include mandatory participations to psychological experiences, which greatly facilitated the recruitment of participants (mean age: 22 years, standard deviation: 3 years). All subjects are French speaking while having different mother tongues: 33 are originally French speaking, 8 Italian, 4 German and 1 Portuguese. According to the self-reports filled by the users, only 20% of participants knew well their teammate.

Each team was welcomed in a room at the department of psychology of the Université de Fribourg-Universität Freiburg, Switzerland, cf. Fig. 1. The main room is located in semi basement and is followed by a smaller room. Thick curtains were kept closed all along the experiments to reduce sunlight variabilities for the video recording. Neon lighting from the ceiling was used instead to have a constant light from session to session. These two rooms were well isolated from external noises, providing thus an appropriate location for data recording.

B. Procedure

Once the participants of a team were introduced to each other, they were then separated in the two rooms and received an introduction on the experiment. They were told that they were taking part to a study focusing on the communication between people by using computer-supported tools, for an overall duration of about an hour long. Each participant first received a questionnaire to evaluate his or her current emotional state by using the Self-Assessment Manikin (SAM) [28], cf. Fig. 2. After this evaluation, participants started to solve individually the survival task for a maximum duration of 10 minutes. Meanwhile, the facilitators of the experiments decided which participant will receive a positive or a negative mood induction according to the self-reported SAM's valence, while properly balancing as far as possible this condition, i.e., having ideally an equal distribution of team members between positive and negative mood's group. Whereas the other participant's mood was targeted to neutral. After the mood induction procedure, which is detailed in the next section, participants engaged in a remote discussion according to the survival task paradigm. Since the construction of the RECOLA corpus was based on a study focusing on emotion perception during remote collaboration, half of the teams used a specific communication setup during the remote collaborative task, as an independent variable. This tool, that we called EmotiBoard, was used to provide a continuous emotional feedback of the teammate during the collaborative task. It aims to study how emotion perception might be facilitated when using a remote conferencing interaction setup. At the end of the experiment, participants completed a form including various self-reports on the task as well as their consent for using, sharing or publishing their data.

C. Mood manipulation

Following the idea of influencing the emotion of participants for collecting emotional data [11], we used a mood induction technique to balance the context of interaction for the collaborative task. The procedure was as follow: while the members of a team were completing the individual task, it was decided whether they will be constituting a positive or a negative group, according to the first SAM self-reports. Participants were considered as a positive group when the trend was positive, i.e., when the mean value of the SAMs' valence was superior to 5, and as a negative group in the other case. For each group, there were always a neutral teammate, i.e., the one with the closest SAM's valence to 5, and either a positive or a negative teammate, i.e., the one with the most distant SAM's valence from 5. The neutral participant watched an animated screen saver (Windows 7 colorful ribbons), to target a neutral mood induction [29], while the other watched either a positive or a negative sequence, according to the self-reported SAM's valence.

Several video clips were tested on a pool of 11 students (6 females, 5 males) to identify which was the best for inducing a positive or a negative mood. The most significant effect for the negative mood induction was obtained on a video clip from the "Sophie's Choice", from William Styron, where a mother has to choose which of her children will be sent in a concentration camp; SAM's mean valence before: 6.6 (2.3), after: 2.2 (1.2). The

TABLE II
MOOD INDUCTION RESULTS VIA THE SAMS

Group	Arousal			Valence		
	Before	After	p	Before	After	p
Pos.	3.7 (1.3)	4.4 (2.0)	0.46	7.4 (1.1)	7.9 (1.0)	0.27
Neu.	3.9 (2.2)	3.6 (1.6)	0.87	6.4 (1.3)	5.6 (1.1)	0.16
Neg.	3.6 (1.9)	4.9 (2.3)	0.17	5.5 (1.4)	3.4 (1.8)	0.01
Neu.	4.1 (1.6)	4.2 (1.7)	0.81	6.8 (1.0)	6.2 (1.4)	0.19

Statistics are given in the following style: [Mean] (standard-deviation); p values were estimated by a non-parametric test, i.e., Kruskal-Wallis.

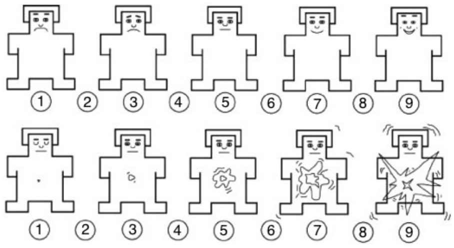

Fig. 2. Self-assessment manikins for valence (up) and arousal (down) [28].

identification of a good video clip for the positive mood induction was more difficult, as it is much harder to influence positively the mood of someone than negatively [30]. From 4 clips, a comedy sketch of the French humorist Gad Elmaleh, ”Le Ski”, was found as providing the strongest positive effect on the pool of subjects; SAM’s mean valence before: 5.1 (1.9), after: 7.0 (1.8). The duration of this video clip was a bit longer than for the negative: 7’15 vs. 4’54, but this difference in duration was necessary to get a significant positive mood induction.

Results of mood induction on the participants of our study are given in Table II. One note that the SAM’s mean valence varies significantly only for the negative mood induction, but participants had already a high emotional valence when they started the experiment. We achieved the goal we set, i.e., balancing the context of interaction, because participants from the positive group had a mood which was significantly different from their (neutral) teammate ($p<0.01$), as well as between participants from the negative group. Furthermore, the number of participants is pretty well balanced for each group, because 12 participants had a negative mood when they started the discussion with their teammate, 24 had a neutral mood and 10 a positive one.

D. EmotiBoard: augmented remote communication

Two large display surfaces served as interactive support for the collaborative task. Skype was used in full-screen for the video-conference and java software was developed to add a continuous feedback of the team-members emotional states. In the EmotiBoard framework, an emoticon showing the emotion of the teammate is superposed on the top-left corner of his or her video. Test facilitators that were acting as wizard-of-oz (WOZ) generated the emotional feedback with a Feeltrace like tool [20], while watching the participant video on an external screen. Emotional ratings of the test facilitators were send to a server and the emoticon seen by the remote participant changed accordingly. The size of the emoticon depended on arousal whereas both smile direction and color of the emoticon depended on valence: a green color and a smile toward the up for a positive valence, and a red color and a smile toward the down for a negative. These three variables (size, color and direction of the smile) were defined in vector graphics to maximize the granularity of the emoticon.

E. Survival task

The task we used for collecting spontaneous collaborative interactions had to be as simple as possible, while ensuring that people would be both motivated and sufficiently involved with regard to their emotions during their communication with their remote teammate. To this end, we choose a task that is frequently used in social psychology for eliciting decision-making processes in small groups: the survival task. This task was originally designed by National Aeronautics and Space Administration (NASA) to train astronauts before the first moon landing [31]. Group discussion is promoted by asking participants to reach consensus on how to survive in a disaster scenario, like a plane crashing in a deserted and hostile area. The group has to rank a number of items (usually 15) according to their importance for crew members to survive. Participants of our study first performed this task individually and then discussed their solution with their teammate until they reach a consensus; mean duration was about 15 minutes.

This kind of task provides a number of advantages for use in the present context. It allows to model the typical process of reaching consensus in a complex decision-making task, where the solution to the problem is not straightforward and may require an intensive group discussion. Since there are high stakes associated with good task performance (i.e. crew survival), this task is likely to trigger emotion, even if used in a laboratory context. It may be possible that some participants wished to hide their real emotion if they showed a negative valence. This may be due to the generally higher social desirability to show positive emotion in a work context.

F. Self-reports of participants

Various self-reports were completed by participants along with the experiment. Participants were first asked to fill a form with information regarding their age, gender and mother tongue. Then they rated their mood 3 times using the SAMs: (i) at the beginning of the task, (ii) after the mood induction and (iii) at the end of the experiment. Participants also evaluated the mood of their teammate at the end of the task using the SAMs, and their emotion using the positive and negative affective schedule (PANAS) [32].

Regarding the social information, participants filled two forms: the team climate inventory (TCI) [33] and the NASA task load index (NASA-TLX) [34]. The TCI covers three domains: (i) communication inside the group, (ii) collaboration between teammates and (iii) satisfaction regarding the accomplished team work. Whereas the NASA-TLX includes questions that are more related to the survival task: (i) task-related scales, (ii) behavior-related scales and (iii) subject-related scales. For the RECOLA corpus, we selected items from both TCI and NASA-TLX questionnaires that were related to the five annotated social dimensions.

IV. MULTI-MODAL RECORDING SETUP

A. Multimodal sensors and softwares

Audio data were captured by unidirectional headset microphones (AKG C520L) and recorded with Audacity software at 44.1kHz, 16bits. An external sound card (Lexicon Omega) was used to split the audio data for being simultaneously used by Skype and Audacity. Two HD webcams (Logitech C270) were used for each participant. The first webcam only captured the video data and the signal was split in two using Splitcam software: one for the Skype video-conference and the other for the visual display used by the WOZ. Whereas the second webcam was used to record both audio, from the built-in omnidirectional microphone, and video with the software provided by the manufacturer; audio was recorded at 48kHz, 16bits and brightness auto adjustments were turned off for video recording. For the physiological data, we used the Biopac MP36 unit and the Biopac Student Lab software to record both EDA and ECG signals at 1kHz.

B. Synchronization

As there was no easy method available to trigger or record synchronization pulses from the webcams, unlike the MP36 unit, we used an inter-correlation maximization technique for synchronizing the audiovisual data. The high quality (HQ) audio signal, which is captured by an external microphone, is sync with the video data by performing an inter-correlation analysis with the low quality (LQ) signal, which is captured by the webcam and already sync with the video frames. The timing of a speech event that was clearly visible in both HQ and LQ signals, like the rise of a plosive, was first manually provided. Then, the signals were normalized and LQ down-sampled to 44.1kHz, to compute their inter-correlation signal from the located event, with a time shift ranging from -10ms to +10ms and a step of 1ms. The delay corresponding to the local maximum of this inter-correlation signal was then used to synchronize the HQ audio data with the video frames. This technique was also used to synchronize the HQ audio data from the dyadic recordings, which allows the study of emotion contagion and interpersonal synchrony between participants of a team. Whereas the physiological signals were synchronized by hardware, using a technique that was originally developed for multi-camera triggering [35]: the right channel of the HQ audio signal recorded the synchronization pulses that were emitted by the MP36 unit, when the recording of the physiological data started. A precision of 1ms can be thus guaranteed for both audiovisual and physiological data synchronization.

C. Data acquisition and compression

Disposable electrodes were fixed on the skin of the participant with medical tape to measure the biosignals; 2 electrodes were fixed at the end of the index and middle finders for EDA recording, and 3 sensors located at the palm of the right hand, and right and left inner ankles served for ECG recording. The headset microphones were placed on the head of the participants and the cameras angle adjusted to have all the face visible on the screen. The audio recording was launched first, followed by the biosignals. According to the condition of the experiment, i.e., using or not the EmotiBoard, the java application was launched. Finally, the video recording was started just before the participants began the collaborative task. All this procedure was simplified with pre-configured scripts. Concerning data compression, we reduced the quality factor (q=25) and changed the frame rate from variable 30Hz to constant 25Hz of the video data by an encoding in mpeg4 using H264 codec.

D. Practical considerations

Despite the efforts we made to have a recording system fully operational, there were various issues during the data collection. First, the two webcams were placed inside a box that was fixed on the table in front of the participants. But this configuration had to be changed quickly because there were a lot of vibrations on the table due to movements of the participant. Webcams were thus moved to be fixed on a chair that was placed close to the table but without touching it. Second, it happened that one of the recording software crashed during the recording, or that the electrodes used for capturing the biosignals were involuntarily removed by the participant. Finally, the last issue we met concerned the order of recording, because we need to have first the audio signal being recorded and then the biosignals, to get the start of the synchronization pulses. All sessions that faced an issue either on audio or video recording were removed from the corpus, whereas those that only had issues with the biosignals were kept. At the end of the acquisition campaign, the RECOLA corpus includes data from 46 participants with audiovisual recordings, from which 35 were recorded with a fully multimodal setting, and 27 of them agreed to share their data.

V. ANNOTATIONS

In order to focus on what we believe to be the most interesting part of the survival task, we kept only the first five minutes of each group recordings. This was due to the participants spending more time discussing their strategies at the beginning of the task, but it also had the additional benefit of limiting the quantity of data to be annotated. From more than 9.5 hours of recordings, we obtained a reduced set of 3.8 hours with audiovisual data from 46 participants, including 2.9 hours of multimodal data. We concentrated on the annotation of both the affective and social behaviors that were produced by participants during their collaboration. Even though only the first 5 minutes of each recording is annotated, the full length of the collaborative task will be made available in the RECOLA corpus.

A. Annotation tool

Different tools were tested for data annotation, like Feeltrace [20] and Gtrace [36]. However, we had to develop our own application because we wanted to use a web-based approach for facilitating the remote annotation of data [30]. Furthermore, we considered that judging two emotional dimensions at the same time, like arousal and valence in Feeltrace, may be too cognitively demanding to reach a high quality on both, especially for a stimuli of 5 minutes. We rather used a setting with one time-continuous annotation for each affective dimension, like in Gtrace. The annotation tool we developed and called ANNEMO is being released in conjunction with this paper[2].

In our setting, the annotator logged in a web-based annotation interface by using a unique identifier and through the Google Chrome web-browser, because it handles various video codecs including H264. The interface is split vertically in two parts: a scrolling list of the audiovisual recordings is given on the left side as an html list, whereas the video and the annotation cursor are displayed one below the other on the right side of the window, cf. Fig. 3. The two affective dimensions (arousal and valence) were annotated separately and time-continuously, using a slider with values ranging from -1 to +1 and a step of 0.01. To avoid delay in data transmission, timestamps were estimated on the local machine as the delay between a slider event, i.e., a change in the annotation, and the video start. Whereas the social dimensions were rated once after having performed the annotation of the affective behaviors, using a 7-Likert scales on the five following dimensions: agreement, dominance, engagement, performance and rapport, which were extracted from various studies in the literature.

B. Annotation guidelines

Each annotator was instructed orally and received instructions with a 4 pages document explaining in details the procedure to follow for the annotation task. This document included a short

[2] Available: http://diuf.unifr.ch/diva/recola/annemo

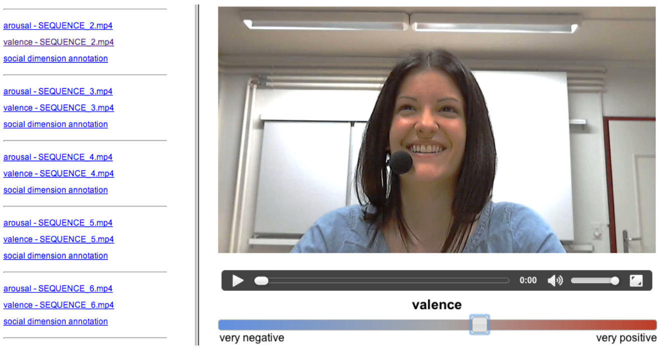

Fig. 3. ANNEMO: web-based annotation of affective and social behaviors.

list of some well identified emotional cues for both arousal and valence, to provide a common introduction on emotions to the annotators, even though they were rather instructed to use their own feeling for the annotation task. Before starting the annotation of the data from the RECOLA corpus, annotators first performed the annotation of two video sequences selected from the SEMAINE corpus [11], to become familiar with the annotation interface while being instructed: we incrusted a screenshot of the annotation cursor used in our tool into the video frames of the SEMAINE sequences, and we changed the position of the cursor into the frames according to the mean value computed from the emotional ratings of the SEMAINE corpus, i.e., arousal and valence. For convenience, annotators were allowed to use a control panel of the video to eventually stop the play of a sequence and restart the annotation at a given instant; new data were re-written over the previous annotation samples. The possibility of step changes in the data is prevented by using post-processing techniques on the continuous annotation, i.e., binning on values provided by each annotator and mean filtering on all annotations of a same sequence.

Concerning the annotation of the social behavior, we provided a list of definition and question to the annotators for each primitive: (i) agreement: does the person seem to agree with his/her partner?, (ii) dominance: does the person appear to be dominant?, (iii) engagement: does the person seem to be engaged?, (iv) performance: does the persons speech appear to be clear and relevant for the task? and (v) rapport: does the person and his/her partner seem to be or could become friends?

C. Annotation data collection

We first hired two assistants to compare empirically the influence of mother tongue on the annotations: one annotator was French speaking whereas the other was not. As one may expect, the annotation provided by the non-French speaking assistant, cf. the thick blue line in Fig. 4, is much less detailed than for those that were made by assistants who had both the verbal and the non-verbal information available for judging the emotion. We therefore recruited only French speaking assistants for a total number of 6 annotators [37]. Ratings were automatically checked with a script all along the collection of the annotations to ensure that: (i) the delay between the first annotation and the video start was less than 5 seconds (ii) the delay between two consecutive annotation samples of a same sequence (i.e., a ”blank”), was no longer than 20 seconds and (iii) the annotation of the social dimensions was performed after the two affective dimensions for each sequence. Sequences that failed one of these criteria were re-annotated.

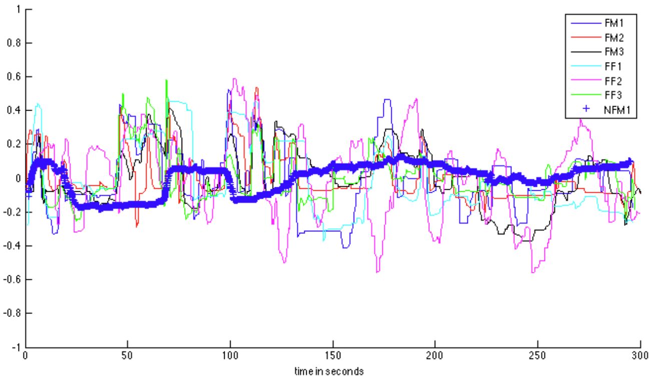

Fig. 4. Annotations of the emotional valence from 7 annotators with zero mean normalization: 3 French males (FM), 3 females (FF) and 1 non-French male (NFM).

D. Post-processing

We performed post-processing of the annotations to reduce unwanted variabilities in the data, such as ”blanks” or jumps due to re-annotation, and to provide a ground truth for the automatic recognition of the annotated behaviors. We used different normalization techniques [38], to study their influence on the inter-rater agreement. In order to deal with issue of missing values in the annotations of the affective behaviors, which was limited to appear with a duration less than 20 seconds, data were interpolated using piecewise cubic interpolation, as it preserves the monotonicity and the shape of the data. Then data from the annotators were binned with a frame rate fixed to match with the one used in the video recording, i.e., a 40ms duration bin. This binning process has the additional benefit to reduce jump effects due to re-annotation of the data.

Because variabilities in human judgments do exist [12], we considered the use of normalization techniques on the annotation data. A local normalization, i.e., for each annotated sequence and for each annotator, was performed on the data using two different techniques: (i) a zero-mean (ZM) to remove an eventual bias in the annotation values, e.g., shifted toward positive or negative values, and (ii) a synchronization to tackle the issue of having different time reactions between the annotators for the annotation of the affective behaviors. The synchronization delay of a given annotation was estimated by minimizing the inter-rater mean squared error (MSE) pair-wisely with the annotations provided by all others annotators, while time-shifting the annotations from -2 to +2 seconds [22]; others range values were tested for time-shifting, but results were almost constant. A mean delay was finally computed from those estimated on the pair-wise combinations to synchronize the annotation. Finally, the ground truth of a sequence is estimated by mean filtering the annotations provided by all the 6 annotators.

E. Analysis

For the analysis of the annotation of the affective behaviors, we computed the percentage of positive frames, the MSE, the mean correlation coefficient and the Cronbach's α, which is an estimate of the internal consistency between annotations; $\alpha > 0.7$ is considered as an acceptable internal consistency and $\alpha > 0.8$ as a good consistency. Results from the raw data show that the internal consistency is acceptable for valence and good for arousal, and the amount of positive frames is balanced for arousal but not for valence, cf. Table III. Whereas the use of normalization techniques

TABLE III

STATISTICS OF THE AFFECTIVE BEHAVIORS AFTER APPLYING LOCAL NORMALIZATION PROCEDURES: NO NORMALIZATION, NORMALIZING TO ZERO MEAN (ZM), COMBINED WITH SYNCHRONIZATION

	Raw		Zero mean		Sync and ZM	
	Aro.	Val.	Aro.	Val.	Aro.	Val.
% pos.	52.1	75.5	54.8	44.1	54.9	44.1
MSE	0.0516	0.0887	0.0294	0.0511	0.0291	0.0510
Corr.	0.435	0.407	0.435	0.407	0.443	0.412
α	0.80	0.74	0.80	0.74	0.80	0.75

TABLE IV

STATISTICS OF THE SOCIAL BEHAVIORS AFTER APPLYING LOCAL NORMALIZATION PROCEDURES: NO NORMALIZATION, AND ZERO MEAN NORMALIZATION

	Raw			Zero mean		
	% neg.	% pos.	κ	% neg.	% pos.	κ
Agreement	19.9	69.2	0.28	30.8	22.5	0.24
Dominance	31.5	31.5	0.20	31.5	38.0	0.23
Engagement	13.8	73.2	0.21	32.6	43.5	0.26
Performance	13.8	67.8	0.31	32.2	39.1	0.31
Rapport	18.8	62.0	0.17	38.4	39.1	0.23

improves the inter-rater agreement measures for both arousal and valence, and provides more well balanced instances for valence.

The analysis of the annotation of the social behaviors includes the percentage of positive and negative sequences, as well as the mean linearly weighted Cohen's κ, which is a measure of inter-reliability; $\kappa > 0.01$ is considered as a slight agreement between annotators and $\kappa > 0.2$ as a fair agreement. Results from the raw data show that a fair agreement is obtained in the annotation of 4 social dimensions on 5, with a balanced amount of positive sequences only for dominance, cf. Table IV. Zero mean normalization improves the Cohen's κ for all dimensions excepted agreement, while balancing the amount of positive and negative sequences for all dimensions.

VI. CONCLUSIONS AND FUTURE WORKS

A new multimodal corpus of spontaneous collaborative and affective interactions in French has been introduced: RECOLA, which is being made available to the research community. 46 users were recorded in dyads during a video conference while completing a task requiring collaboration. Recordings include multimodal data (audio, video, ECG and EDA) that were all recorded continuously and synchronously. Various self-reports completed by participants along the task are supplied for both emotional and social behaviors, which were also annotated by 6 French speaking annotators on all recorded sequences. The analysis of the annotations shows a good inter-annotator agreement rate for the affective dimensions, and a fairly good one for the social dimensions, with a balanced distribution of instances when mean centering is applied. However, the relevance of this technique for emotion recognition may be questionable, because the balance of instances can change significantly.

We plan in the future to analyze in more details the annotations provided by both the participants and the external annotators. We will also take into consideration the reaction time delays of each annotator (using different stimuli) in the corpus, and study their influence on emotion recognition from speech using features from the state-of-the-art [39], [40], which will provide recognition baselines to the corpus. Once the relevant models for multimodal emotion recognition will be identified through testing on the RECOLA corpus, they will be integrated in the EmotiBoard application to replace the WOZ human judges by real-time and automated emotion recognizers. This tool will be finally used to measure the impact of such automated emotional feedback on teamwork quality and efficiency.

VII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of Alessia Bernasconi, Clémentine Duc and Cindy Pasquier, from the department of psychology of the Université de Fribourg-Universität Freiburg, who performed all the data collection and played the WOZ roles, and the contribution of Ilya Boyandin, from the department of informatics of the same university, who developed the web-based annotation tool.

REFERENCES

[1] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag German audio-visual emotional speech database," in *Int. Conf. on Multimedia and Expo*, Hannover, Germany, 2008, pp. 865–868.

[2] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 2, pp. 1743–1759, 2009.

[3] C.-C. Lee, A. Katsamanis, M. Black, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Computing vocal entrainment: a signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech and Language*, 2012.

[4] M. Stewart, C. McAdam, M. Ota, S. Peppe, and J. Cleland, "Emotional recognition in autism spectrum conditions from voices and faces," *Autism*, 2012.

[5] E. Douglas-Cowie and al., "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," in *2nd Int. Conf. on Affective Computing and Intelligent Interaction*, Lisbon, Portugal, 2007, pp. 488–500.

[6] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: a review," *Visual and multimodal analysis of human spontaneous behaviour, Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[7] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: current trends and future directions," *Image and Vision Computing*, 2012.

[8] J. Kim, "Robust speech recognition and understanding," *Vienna: I-Tech Education and Publishing, 2007, Bimodal Emotion Recognition using Speech and Physiological Changes*, pp. 265–280, 2007.

[9] M.-Z. Poh, D. McDuff, and R. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 253–263, 2010.

[10] J. Forgas, "The influence of mood on perceptions of social interactions," *J. of Experimental Social Psychology*, vol. 20, no. 6, pp. 497–513, 1984.

[11] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[12] A. Tversky, "Intransitivity of preferences," *Psychological Review*, vol. 76, pp. 31–48, 1969.

[13] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: towards a new generation of databases," *Speech Communication*, vol. 40, pp. 33–60, 2003.

[14] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony : A survey of evaluation methods across disciplines," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 349–365, 2012.

[15] T. Banziger and K. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: the GEMEP corpus," in *A. Paiva, R. Prada, and R. W. Picard, Int. Conf. on Affective Computing and Intelligent Interaction*, Lisbon, Portugal, 2007, pp. 476–487.

[16] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge," *Sensing Emotion and Affect - Facing Realism in Speech Processing, Speech Communication*, pp. 1062–1087, 2011.

[17] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[18] P. Ekman, *Emotion in the human face*. Cambridge, UK: Cambridge University Press, 1982.

[19] J. Russel, "A circumplex model of affect," *Journal of Personality and Social Psychology*, pp. 1161–1178, 1980.

[20] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': an instrument for recording perceived emotion in real time," in *Douglas-Cowie, E., Cowie, R., Schröder, M., Eds.: ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, Belfast, UK, 2000, pp. 19–24.

[21] D. Grandjean, D. Sander, and K. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and Cognition*, vol. 17, pp. 484–495, 2008.

[22] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[23] S. Kitayama, M. Karasawa, and B. Mesquita, "Cultural affordances and emotional experience: Socially engaging and disengaging emotions in Japan and the United States," *Journal of Personality and Social Psychology*, vol. 91, no. 5, pp. 890–903, 2006.

[24] J. Fontaine, E. Roesch, and P. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 2, pp. 1050–1057, 2007.

[25] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, 2012.

[26] S. Gilroy, M. O. Cavazza, and V. Vervondel, "Evaluating multimodal affective fusion using physiological signals," in *Int. Conf. on Intelligent User Interfaces*, Palo Alto (CA), USA, 2011, pp. 53–62.

[27] J. Kelly and S. Barsade, "Mood and emotions in small groups and work teams," *Organizational Behavior and Human Decision Processes*, vol. 86, no. 1, pp. 99–130, 2001.

[28] M. Bradley and P. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *J. of Behavior Therapy and Experimental Pscychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[29] J. Rottenberg, R. D. Ray, and J. Gross, "Emotion elicitation using films," *The handbook of emotion elicitation and assessment*, pp. 9–28, 2007.

[30] A. Göritz and K. Moser, "Web-based mood induction," *Cognition and Emotion*, vol. 20, no. 6, pp. 887–896, 2006.

[31] J. Hall and W. Watson, "The effects of a normative intervention on group decision-making performance," *Human Relations*, vol. 23, no. 4, pp. 299–317, 1970.

[32] J. R. Crawford and J. D. Henry, "The positive and negative affect schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample," *British J. of Clinical Psychology*, vol. 43, no. 3, pp. 245–265, 2004.

[33] N. Anderson and M. West, "Measuring climate for work group innovation: development and validation of the team climate inventory," *J. of Organizational Behavior*, vol. 19, no. 3, pp. 235–258, 1998.

[34] S. Hart and L. Staveland, "Development of NASA-TLX (Task Load Index): results of empirical and theoretical research," *Human Mental Workload*, pp. 139–183, 1988.

[35] J. Lichtenauer, J. Shen, M. Valstar, and M. Pantic, "Cost-effective solution to synchronised audio-visual data capture using multiple sensors," in *Int. Conf. on Advanced Video and Signal Based Surveillance*, Boston (MA), USA, 2010, pp. 324–329.

[36] R. Cowie and M. Sawey. (2011) Gtrace – General trace program from Queen's, Belfast. [Online]. Available: https://sites.google.com/site/roddycowie/work-resources

[37] R. Cowie and G. McKeown. (2010) Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme. [Online]. Available: http://www.semaine-project.eu/

[38] M. Nicolaou, H. Gunes, and M. Pantic, "Automatic segmentation of spontaneous data using dimensional labels from multiple coders," in *LREC Int. Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing*, Istambul, Turkey, 2010, pp. 43–48.

[39] F. Eyben, M. Wöllmer, and B. Schuller, "The Munich versatile and fast open-source audio feature extractor," in *ACM Multimedia (MM)*, Florence, Italy, 2010, pp. 1459–1462.

[40] F. Ringeval, M. Chetouani, and B. Schuller, "Novel metrics of speech rhythm for the assessment of emotion," in *Interspeech*, Portland (OR), USA, 2008, pp. 2763–2766.