THE SEMAINE CORPUS OF EMOTIONALLY COLOURED CHARACTER INTERACTIONS

Gary McKeown[‡], *Michel F. Valstar*[†], *Roderick Cowie*[‡], *Maja Pantic*[†§]

Twente University, EEMCS[§]
Imperial College London, Department of Computing[†]
Queen's University Befast, Department of Psychology[‡]
michel.valstar@imperial.ac.uk, g.mckeown@qub.ac.uk, R.Cowie@qub.ac.uk, m.pantic@imperial.ac.uk

ABSTRACT

We have recorded a new corpus of emotionally coloured conversations. Users were recorded while holding conversations with an operator who adopts in sequence four roles designed to evoke emotional reactions. The operator and the user are seated in separate rooms; they see each other through teleprompter screens, and hear each other through speakers. To allow high quality recording, they are recorded by five high-resolution, high framerate cameras, and by four microphones. All sensor information is recorded synchronously, with an accuracy of 25 μs. In total, we have recorded 20 participants, for a total of 100 character conversational and 50 non-conversational recordings of approximately 5 minutes each. All recorded conversations have been fully transcribed and annotated for five affective dimensions and partially annotated for 27 other dimensions. The corpus has been made available to the scientific community through a web-accessible database.

Keywords— Affective Computing, Emotional Corpora, Multimedia Databases, Emotional Annotations, Sensitive Artificial Listener

1. INTRODUCTION

A key goal for the next generation of Human Computer Interaction devices is "really natural language processing" [1], which allows human users to speak to machines just as they would speak to another person. Achieving that goal depends on finding ways to handle the non-verbal signals that are part and parcel of human conversation. These include signals associated with conversation management (e.g. backchannelling and turn-taking) and the emotional colouring of everyday communication (outbursts of "pure" emotion are a different problem).

Recently a scenario in which other factors are simplified enough to let those issues be addressed systematically has been developed [2]. The scenario is called the Sensitive Artificial Listener, SAL for short. Recording humans interacting in a SAL environment is a fundamental step towards assessing machine-human interactions within this scenario. It involves a user interacting with "characters" whose responses are stock phrases keyed to the user's emotional state rather than the content of what he/she says. The model is a style of interaction observed in chat shows and parties, which aroused interest because it seems possible that a machine with some basic emotional and conversational competence could sustain such a conversation, without needing to be competent with fluent speech and language understanding. This is important, because neither fluent speech generation nor automatic natural language understanding are currently possible [3]. If a machine could maintain a SAL conversation, it would provide an ideal testbed for research concerned with developing that emotional and conversational competence.

Early research modelled the scenario in which users interact with a human operator whose responses were restricted to phrases from a script [4], with preset rules determining when to use which phrase. This version allowed sustained interaction with four "characters", each with a distinct emotional style, and a conversational goal of shifting the user towards that state. They are Prudence, who is even-tempered and sensible; Poppy, who is happy and outgoing; Spike, who is angry and confrontational; and Obadiah, who is depressive. Each is defined by a script of utterances in keeping with the character, and selection rules which chose an utterance calculated to push the user towards the character's state. In spite of its simplicity, that structure turns out to be capable of engaging users in a lively exchange lasting for about half an hour, and involving a rich variety of conversational behaviours.

Previous recordings have been collected using similar SAL scenarios. The unique feature of these SAL scenarios is that they show users interacting in an emotionally coloured conversational style with a machine-like agent. That makes them invaluable to research aimed towards that kind of interaction. This paper describes a key addition to the range of SAL scenarios. The additions include users being physically separated but retaining the possibility of eye-contact and face to face communication, and not using scripts; with the goal of eliciting natural conversation-related behaviours.

A number of publicly available databases that have continuous affect labelling and are thus of use to the Social Signal Processing and Human Behaviour Analysis communities already exist. The Montreal Affective Voices (MAV) database consists of 90 outbursts of affective vocalisations [5]. The vocalisations were produced on command by five male and five female actors. Each vocalisation was annotated by a single value on a scale of $[0, 100]$ by 30 naive raters for the dimensions valence, arousal,

978-1-4244-7493-6/10/$26.00 ©2010 IEEE

Table 1. Overview of databases with continuous labelling of emotionally coloured content.

Database	Participants	Duration	Sensors	Video Bandwidth	Audio Bandwidth	Web-based	Searchable
Montreal Affective Voices [5]	10	unknown	1	No video	44.1 kHz	No	No
Vera am Mittag db [6]	20	12:00:00	3	352x288 pixels @ 25Hz	16kHz	No	No
SAL [4]	4	4:11:00	2	352x288 pixels @ 25 Hz	20kHz	Yes	No
SEMAINE	20	6:30:41	9	580x780 pixels @ 49.979 Hz	48kHz	Yes	Yes

and 8 other affective states.

The Vera am Mittag (VAM) database consists of 12 hours of televised talk-show episodes [7]. The audio part of these data have been annotated for valence, arousal, and dominance on a continuous time, continuous value scale by 17 raters. The video element was annotated for a subset of the recordings. For 20 participants of the talkshow the frames in which a face was present were annotated by 8-34 raters for the dimensions valence, arousal and dominance. In addition, the six basic emotions were rated on a continuous time and value scale.

The SAL database consists of recordings made with the SAL technique [2]. Four participants were recorded by a camcorder while having a conversation with an experimenter who pretended to be an automatic agent (see section 2 below). It was fully annotated on a continuous time, continuous value scale for the dimensions valence and arousal by four raters.

Table 1 lists the properties of the existing databases. From the table a number of shortcomings become clear. First of all, all existing databases have relatively low spatial and temporal video resolution. These resolutions are too low for sensitive automatic affect recognition. Also, the number of participants of the MAV and SAL databases are too low to draw any general conclusions. The VAM corpus does have a significant number of participants. However, the VAM corpus has another major drawback: because it is an edited television programme, the camera views and microphone sources change very frequently. Lastly, although the SAL database is available through an online interface, none of the existing databases provide the facility to search through the corpus, allowing researchers to select and download the dataset they require. These are all issues that we have sought to overcome in the creation of the SEMAINE corpus of emotionally coloured character interactions.

2. RECORDING EMOTIONALLY COLOURED CONVERSATIONS

Previous SAL recordings have used a “Wizard of Oz” like scenario where an operator navigates a complex script and decides what the “agent” should say next (in older versions, the operator speaks the phrase; in recent versions, he or she selects the utterance and the system generates it). These recordings are emotionally interesting, but because the operator is looking at the script, not the user, the conversational elements are very limited. The “Solid SAL” scenario was developed to overcome that problem. The operator is required to become thoroughly familiar with the SAL characters, and speaks as they would. Several factors help to maintain the impression of machine interaction. The scenario is explained to the user; and the interaction takes place at a distance. Participants see each other’s face in a teleprompter screen, and hear through loudspeakers; and probably most important, it quickly becomes clear that the operator is not obeying the rules of a normal human-human interaction as they adopt the character roles. Pilot studies confirmed that interactants show a rich set of conversation-related behaviours, in particular back-channelling and turn-taking, along with emotional behaviours comparable to previous versions. Formal recordings were therefore collected.

2.1. Experimental Procedure

The experiment was designed as the first in a sequence of situations that mimic the SAL interaction. This first experimental situation is known as Solid SAL. The setup is designed to capture a human-human conversation between two interacting participants with many of the constraints of a machine agent-human conversation (e.g. conversation with a disembodied, two-dimensional screen based image) but retaining an ability to evoke natural conversation-related behaviours.

Participants were recruited from undergraduate and postgraduate students at Queen’s University Belfast and our research team. Participants read a written participant information sheet and provide written consent before a verbal explanation of the SEMAINE project goals, what they are required to do in the experiment and a description of each character. A typical session lasts about twenty minutes with an approximate interaction time of five minutes per character. The actual duration of character interactions varies depending individual interactions. Participants are told to request a change of character when they get bored, annoyed or feel they have nothing more to say to the character. The operator can also request to change character if enough time has passed with a character or a conversation has a natural conclusion.

Participants are next brought to the recording studio, where they sit down and put on their head microphone. The operator takes her/his place in a separate room and recording starts. The operator recites a brief introduction to Solid SAL script and asks the user which character they wish to speak to, after which the conversational interaction begins. The operator is required to act each of the four characters in turn, but in no particular order and usually the user decides the order.

2.2. Interaction Scenario

The operator plays the role of each of the SAL characters in turn in as natural a manner as possible without recourse to a script and maintaining a natural style of conversation within the

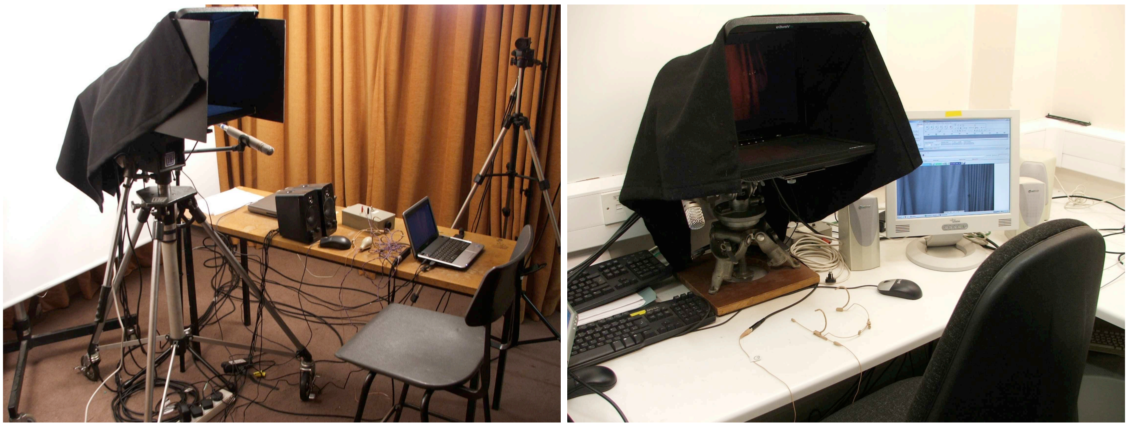

Fig. 1. Images of the recording setup for both the user and operator Rooms.

constraints of the role of the character. The user interacts in conversation with each of these characters in as natural a manner as possible. Pilot attempts using scripts and learned repertoires resulted in conversation that was stunted and not natural so a more free form approach was adopted instructing operators to play the roles of the characters. Users are encouraged to interact with the characters as naturally as possible. There is a single explicit constraint: users are not permitted to ask questions to the characters. When questions are asked by users they are reminded by the operator that the SAL characters cannot answer questions. In some situations the operators do not comply with these rules and answer questions and incorporate knowledge from the conversation. However the operators are instructed that the most important aspect of their task is to create a natural style of conversation; strict adherence to the rules of a SAL engagement was secondary to a conversational style that would produce a rich set of conversation-related behaviours and therefore transgressions occasionally occur. Once all four of the characters have interacted with the user the operator brings the recording session to a close.

3. SYNCHRONISED MULTI-SENSOR RECORDING SETUP

The database is created with two distinct goals in mind. The first is the analysis of this type of interaction by cognitive scientists. This means that the recordings should be suitable for use by human raters, who intend to analyse both the auditive and the visual communication channels. Secondly, the data is intended to be used for the creation of machines that can interact with humans by learning how to recognise social signals. The goal for the machines is to use both the auditive and the visual modalities.

Sensors. Video is recorded at 49.979 frames per second at a spatial resolution of 780 x 580 pixels, while audio is recorded at 48 kHz with 24 bits per sample. Both the user and the operator are recorded from the front by both a greyscale camera and a colour camera. In addition, the user is recorded by a greyscale camera positioned on one side of the user to capture a profile view of their face. An example of the output of all five cameras is shown in Fig. 2.

The reason for using both a colour and a greyscale camera is directly related to the two target audiences. A colour camera needs to interpolate the information from four sensitive chip elements to generate a single pixel, while the greyscale camera needs only a single sensitive chip element. The greyscale camera will therefore generate a sharper image. Machine vision methods usually prefer a sharp greyscale image over a blurrier colour image. For humans however, it is more informative to use the colour image [8].

To record what the user and the operator are saying, we use two microphones per person: the first is placed on a table in front of the user/operator, and the second is worn on the head by the user/operator. This results in a total of four microphones and thus four recorded channels. The wearable microphone is the main source for capturing the speech and other vocalisations made by the user/operator, while the room microphones are used to model background noise.

Environment. The user and operator are located in separate rooms. They can hear each other over a set of speakers, which output the audio recorded by the wearable microphone of their conversational partner. They can see each other on teleprompters. The frontal cameras are placed behind the semi-reflecting mirror. This way, the user and operator have the sensation that they look each other in the eye. This proved to be very important, as a pilot test where the cameras were placed on top of a screen did not evoke the sensation of eye-contact, which is essential in human-human communication. Professional lighting was used to ensure an even illumination of the faces.

Synchronisation. In order to do multi-sensory fusion analysis of the recordings, it is extremely important to make sure that all sensor data is recorded with the maximum synchronisation possible. To do so, we used a system developed by Lichtenauer et al. [9]. This system uses the trigger of a single camera to accurately control when all cameras capture a frame. This ensures

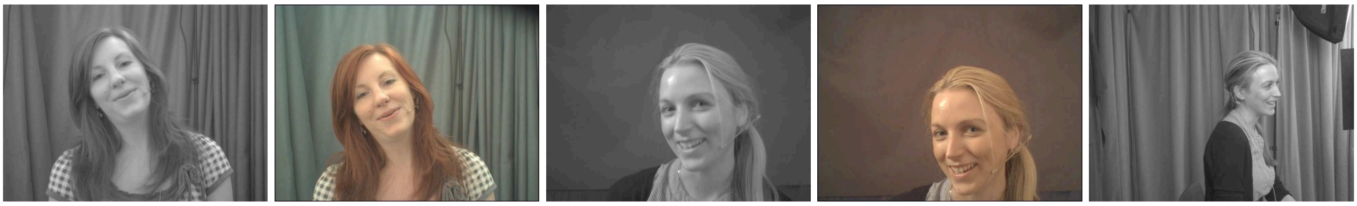

Fig. 2. Frames grabbed at a single moment in time from all five video streams. The operator has HumanID 7, and the user has HumanID 14. Shown is the 3214th frame of the 19th recording.

all cameras record every frame at almost exactly the same time. The same trigger was presented to the audio board and recorded as an audio signal together with the four microphone signals. This allowed us to synchronise the audio and the video sensor data with an accuracy of 25 microseconds.

Data compression The amount of raw data generated by the visual sensors is very high: 25 recordings, lasting on average 30 minutes, recorded at 49.979 frames/second at a temporal resolution of 780*580 pixels with 8 bits per pixel for 5 cameras, results in 4.627 TeraByte. This is impractical to deal with: it would be too costly to store and it would take too long to download over the internet. Therefore, the data has been compressed using the H.264 codec and stored in an avi container. The video was compressed to 440 kbit/s for the greyscale video and to 500 kbit/s for the colour video. The recorded audio was stored without compression, because the total size of the audio signal was small enough.

4. ANNOTATIONS

Trace style continuous ratings were made on five core dimensions for all recordings where annotators are instructed to provide ratings for their overall sense of where an individual should be placed along a given dimension [4]. These dimensions are those that psychological evidence suggests are best suited to capture affective colouring in general [10]. They are Valence, Activation, Power, Anticipation/Expectation with the addition of Overall Emotional Intensity. Valence and Activation are the most widely recognised affective dimensions, which is why they provide the basis for the four Solid SAL characters. Valence labels whether the annotator on balance feels positive or negative about the things, people, or situations at the focus of their emotional state. Activation is the individual's overall inclination to be active or inactive. Activation may include mental as well as physical activity, preparedness to act as well as overt activity, alertness as well as output, directed or undirected thought. Power and Expectation along with Valence and Activation make up the four most important dimensions of emotion according to Fontaine, Scherer, Roesch and Ellsworth [10]. To these four dimensions we add a measure of the overall level of emotional intensity is the simplest description of an emotional state. Roughly speaking it is about how far the person is from a state of pure, cool rationality, whatever the direction.

Once raters have annotated the five core dimensions they then choose an additional four out of a possible 27 optional rating dimensions (four was considered feasible given resource limitations). Occasionally more than four are chosen if four is not considered sufficient. These dimensions are only annotated where a rater feels that there has been at least one obvious instance of the relevant phenomenon within the section being rated. The additional dimensions are divided into four categories, as shown in Table 2. These categories are derived from four different backgrounds, basic emotions, epistemic states, interaction process analysis and validity.

There is a widespread belief that basic emotions are important points of reference even in material that involves emotional colouring rather than prototypical emotional episodes. Hence the commonest list of basic emotion terms, Ekman's "big six" [11], is included. To this is added a category that would not otherwise be represented, amusement. These are labeled where a clip contains a relatively clear-cut example of the basic emotion. The included labels for *Emotion* are shown in column one of Table 2. These are fundamentally categories that can be located within the more traditional emotional space derived from the dimensions such as Valence, Activation, Power, Anticipation/Expectation [10]. However for annotation purposes we use the same dimensional rating tools for the analysis of these basic emotion categories. These annotations are best thought of as measures of the intensity of a given category and not an dimensions that can placed in more conventional emotional spaces. They can be converted into more traditional binary labels of happiness by setting an acceptable threshold level of intensity.

Epistemic states were highlighted by [12], and have aroused a lot of interest in the machine perception community. They are relatively self-explanatory. As before, these are labeled where the clip contains a relatively clear-cut example of epistemic state. The guidelines for selection of certainty, for example, suggest inclusion if there is an episode where the fact that someone is certain about something stands out, then it warrants inclusion; but not if they simply feel that the person probably has no active doubts about something (e.g. that the sun will rise tomorrow). The included labels for *Epistemic states* are shown in the second column of Table 2.

A further set of labels of particular use in dialogue management is a subset derived from the system of categories used in Interaction Process Analysis [13]). Although these labels will not be used to recreate the method of interaction process analysis, their inclusion should allow a partial reconstruction of the

Table 2. Additional Annotation Dimensions by Category (numbers represent the raw number of annotations for each dimension at time of print).

Basic Emotions		Epistemic States		Interaction Process Analysis		Validity	
10	Anger	23	(not) certain	9	Shows solidarity	11	Breakdown of engagement
2	Disgust	79	(dis) agreement	15	Shows Antagonism	0	Anomalous Simulation
82	Amusement	22	(un) interested	12	Shows tension	19	Marked sociable Concealment
27	Happiness	39	(not) at ease	14	Releases tension	5	Marked sociable simulation
21	Sadness	41	(not) thoughtful	6	Makes suggestion		
10	Contempt	9	(not) concentrating	2	Asks for suggestion		
				42	Gives Opinion		
				3	Asks for opinion		
				72	Gives information		
				3	Asks for information		

method. The included subset of interaction process analysis labels are shown in Table 2.

The final set of labels assess validity to some extent. The aim is to highlight areas of the data where the observed effect and corresponding label differ from those on which the system should be trained. Breakdown of engagement seeks to identify periods where one or both participants are not engaging with the interaction. Anomalous simulation seeks to identify periods where there is a level of acting that suggests the material is likely to be structurally unlike anything that would happen in a social encounter. Marked sociable concealment is concerned with periods when it seems that a person is feeling a definite emotion, but is making an effort not to show it. Marked sociable simulation is concerned with periods when it seems that a person is trying to convey a particular emotional or emotion-related state without really feeling it.

These dimensions are all recorded as continuous trace style ratings with values between between -1 (minimum presence or one extreme of a dimension) and +1 (maximum presence or the other extreme of a dimension). Utilities are provided with the database for converting these continuous labels into categorical labels at desired thresholds. Up to four raters are used in annotating the data, levels of inter rater reliability differ depending on the dimension. A subset of the clips have been rated by four raters, most have been rated by less. However, annotations are on-going at the time of print. Most of the annotations relate to the user clips due to a greater utility of user annotations in the SEMAINE but a small subset of operator clips have also been annotated.

Besides these dimensional annotations, transcripts are provided for each of the clips containing text of the spoken interaction and minimal non-systematic non-verbal descriptions. Linguistic analysis of these texts is conducted using the Linguistic Inquiry and Word Count (LIWC, [14]) program providing an additional 68 linguistic dimensions for the global conversation, user speech and operator speech for each clip. The most important of these are the emotion related dimensions, Affective processes, Positive emotion, Negative emotion, Anxiety, Anger, Sadness which provide somewhat overlapping "bag of words" style analysis of the emotional content in the verbal communication in each clip. These analyses count the number of words in a given category that occur within the conversations. For example, the words "love," "nice" and "sweet" occurring in the conversation would increase the positive emotion score by three similarly "hate," "kill" and "annoyed" would increase the anger score and "crying," "grief" and "sad" would increase the sadness score. These can be compared against norms for spoken language are provided with LIWC that are derived from transcripts of 2,014 real world unstructured talking situations.

5. WEB ACCESSIBLE DATABASE

The SEMAINE Solid-SAL dataset is made freely available to the research community. It is available through a web-accessible interface with url http://www.semaine-db.eu The available dataset consists of 25 recordings, featuring 21 participants. Four of these participants play the role of the operator in the sessions, but they also appear in the user role in some of the interactions. At time of recording, the youngest participant was 22, the oldest 60, and the average age is 32.8 years old (std. 11.9), 38% are male, and the participants come from 8 different countries. Unfortunately, all but one participants come from a Caucasian background, making the dataset ethnically biased.

Organisation. Within the database, the data is organised in units that we call a *session*. A session is part of a recording, in which the user speaks with a single character. There are also two extra special sessions per recording, to wit, the *recording_start* and *recording_end* sessions. These sessions include footage of the user/operator preparing to do the experiment, or ending the experiment. Although these sessions do not show the desired user/character interaction, they may still be useful for training algorithms that do not need interaction, such as the facial point detectors of detectors which sense the presence of a user.

Each session has 11 sensor data files associated with it. We call these database entries *tracks*. Nine of these are the five camera recordings and the four microphone recordings (see Section 3). In addition, each session has two lower-quality audio-visual tracks, one showing the frontal colour recording of the user, and the other showing the frontal colour recording of the operator. Both low-quality recordings have audio from the operator and the user. The use of these low-quality recordings lies in the fact that they have both audio and video information, which makes them useful for the annotation of the conversation by human raters. To allow annotators to focus on only one person talking, we stored the user audio in the left audio channel, and the operator audio in the right audio channel. Because most media players have a *balance* slider, a human rater can easily choose who to listen to. The low-quality audio-visual tracks are also fairly small which makes them more convenient for download.

In our database, all annotation files (annotations) are associated with a track. It is possible that a single annotation belongs to multiple tracks: for instance, the affective state of the user is associated with all tracks that feature the user. Other annotations can be associated with only a single track.

In the web-accessible database interface, sessions, tracks, and annotations are displayed conveniently in a tree-like structure. One can click on the triangles in front of tree nodes to view all branches. Apart from the tracks and annotations, each session also shows information of the people that are present in the associated recording. This information about the persons shown is anonymous: it is impossible to retrieve a name of the subject from the database. In fact, this information is not even contained in the database.

Search. To allow researchers to conveniently find the data they require, we have implemented extensive database search options. Searching the database can be done either by using regular expressions or by selecting elements to search for in a tree-structured form. The regular expression search is mainly intended for people who work with the database on a day to day basis and who know the search options by heart.

Search criteria can use characteristics of sessions, subjects, tracks, and annotations. It is possible to search by user gender, age, and nationality, by session character, by active AUs, and many many more. Once a search is concluded, the user can inspect the properties of the returned sessions, tracks, and annotations. It is also possible to watch a preview of all the returned video tracks.

6. CONCLUSION

The Solid SAL component of the SEMAINE corpus deliberately addresses a necessary data requirement to reach the goal of next generation of Human Computer Interaction. It does this utilising a scenario that elicits important non-verbal signal components of conversation, conversation management such as backchannelling and turn-taking and the emotional colouring of a conversation in an interactive and dynamic setting. The corpus and annotations are now publicly available and will further advances towards the goal of producing "really natural language processing" [1].

7. ACKNOWLEDGMENTS

This work has been funded in part by the European Community's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE). The work of Michel Valstar and Maja Pantic is further funded in part by the European Community's 7th Framework Programme [FP7/20072013] under the grant agreement no 231287 (SSPNet).

8. REFERENCES

[1] R Cowie and M Schröder, "Piecing together the emotion jigsaw," *Machine Learning for Multimodal Interaction*, pp. 305–317, Jan 2005.

[2] E Douglas-Cowie, R Cowie, C Cox, N. Amir, and D Heylen, "The sensitive artificial listener: an induction technique for generating emotionally coloured conversation," in *Proc. Workshop Corpora for Research on Emotion and Affect*, 2008.

[3] M Schröder, "Expressive speech synthesis: Past, present, and possible futures," *Affective Information Processing*, pp. 111–126, May 2009.

[4] E Douglas-Cowie et al., "The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data," *Lecture Notes in Computer Science*, vol. 4738, pp. 488–501, Jan 2007.

[5] P. Belin, S. Fillion-Bilodeau, and F. Gosselin, "The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing," *Behavior Research Methods*, vol. 40, pp. 531–539, 2008.

[6] M Grimm, K Kroschel, and S Narayanan, "The vera am mittag german audio-visual emotional speech database," *Multimedia and Expo, 2008 IEEE International Conference on*, pp. 865–868, 2008.

[7] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*, 23 2008-April 26 2008, pp. 865–868.

[8] Valdez and Mehrabian, "Effects of color on emotions," *Journal of Experimental Psychology-General*, vol. 123, pp. 394–408, 1994.

[9] J.F Lichtenauer, J Shen, M.F Valstar, and M Pantic, "Cost-effective solution to synchronised audio-visual data capture using multiple sensors," in *Proc. IEEE Int'l Conf' Advanced Video and Signal Based Surveillance*, Nov 2010, pp. 324–329.

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

[11] P Ekman, "Expression and the nature of emotion," *Approaches To Emotion*, pp. 1–25, Jan 1984.

[12] S. Baron-Cohen, O. Golan, S. Wheelwright, and J. J. Hill, "Mind reading: the interactive guide to emotions," 2004.

[13] F Bales, "Interaction process analysis: a method for the study of groups," pp. 203–245, Jan 1951.

[14] J.W Pennebaker, C.K Chung, M Ireland, A Gonzales, and R.J Booth, "The development and psychometric properties of liwc2007," pp. 1–22, Oct 2007.