The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schröder, M. (2012). The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing*, *3*(1), 5-17. https://doi.org/10.1109/T-AFFC.2011.20

Published in:
IEEE Transactions on Affective Computing

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent

Gary McKeown, Michel Valstar, *Member*, *IEEE*, Roddy Cowie, *Member*, *IEEE*, Maja Pantic, *Fellow*, *IEEE*, and Marc Schröder

Abstract—SEMAINE has created a large audiovisual database as a part of an iterative approach to building Sensitive Artificial Listener (SAL) agents that can engage a person in a sustained, emotionally colored conversation. Data used to build the agents came from interactions between users and an "operator" simulating a SAL agent, in different configurations: Solid SAL (designed so that operators displayed an appropriate nonverbal behavior) and Semi-automatic SAL (designed so that users' experience approximated interacting with a machine). We then recorded user interactions with the developed system, Automatic SAL, comparing the most communicatively competent version to versions with reduced nonverbal skills. High quality recording was provided by five high-resolution, high-framerate cameras, and four microphones, recorded synchronously. Recordings total 150 participants, for a total of 959 conversations with individual SAL characters, lasting approximately 5 minutes each. Solid SAL recordings are transcribed and extensively annotated: 6-8 raters per clip traced five affective dimensions and 27 associated categories. Other scenarios are labeled on the same pattern, but less fully. Additional information includes FACS annotation on selected extracts, identification of laughs, nods, and shakes, and measures of user engagement with the automatic system. The material is available through a web-accessible database.

Index Terms—Emotional corpora, affective annotation, affective computing, social signal processing.

1 Introduction

One of the natural long-term goals in affective computing is to develop systems that can engage a human being in a face-to-face conversation which is fluent, sustained, and emotionally colored [1], [2], [3]. This paper describes one of the first databases to be developed with that goal in mind, as part of a project called Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression (SEMAINE). The database includes high-quality, multimodal recordings showing a range of related interactions. At one end of the range are recordings showing pairs of people engaged in emotionally colored conversations. At the other end are recordings of individuals interacting with an automatic system that simulates one of the parties in the human-human recordings. The humans show a range of responses to the system's efforts, from lively interaction to irritated disengagement. Innovative techniques are used to label the material, much of it in considerable depth.

The database provides resources for work on diverse problems associated with fluent interaction: describing relevant processes, particularly nonverbal processes, as cognitive scientists do; training systems to recognize emotion-related states as they appear in conversation, particularly states that emerge in response to a machine attempting to converse; and finding ways to label these states. Critically, it offers support for work that aims not just to describe or to build components, but to build systems that actually have face-to-face emotional interactions with human beings because they deal with a scenario developed specifically to make that possible.

1.1 The Motivation for the Database

It has only gradually become apparent that affective computing might need data on something as specific as fluent, sustained, emotionally colored conversation: But, there are growing indications that building a system to function in a particular context requires data from contexts that are quite similar [4], [5]. Emotion is inherently interactive, and so the states that arise in a given situation, and the signs associated with them, are likely to be a function of the interactions that take place there [6].

Databases dedicated to emotionally colored face-to-face conversations are in short supply. Many databases serve related functions, but very few serve that particular one. The AMI meeting database shows realistic sustained interactions, but they are not rich in emotion [7]. Various databases

- *G. McKeown and R. Cowie are with the School of Psychology, Queen's University Belfast, David Keir Building, Belfast BT7 1NN, Northern Ireland, United Kingdom. E-mail: {g.mckeown, r.cowie}@qub.ac.uk.*
- *M. Valstar is with the Department of Computing, Imperial College London, 180 Queen's Gate, South Kensington Campus, London SW7 2AZ, United Kingdom. E-mail: Michel.Valstar@imperial.ac.uk.*
- *M. Pantic is with the Department of Computing, Imperial College London, 180 Queen's Gate, South Kensington Campus, London SW7 2AZ, United Kingdom, and the University of Twente, The Netherlands. E-mail: m.pantic@imperial.ac.uk.*
- *M. Schröder is with DFKI GmbH, Language Technology Lab, Campus D3 2, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany. E-mail: marc.schroeder@dfki.de.*

Manuscript received 30 Nov. 2010; revised 4 May 2011; accepted 15 June 2011; published online 12 July 2011.
Recommended for acceptance by B. Schuller.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFCSI-2010-11-0119.
Digital Object Identifier no. 10.1109/T-AFFC.2011.20.

of acted material [8], [9], [10], [11] show how people express emotion deliberately, but not how it arises spontaneously in the course of an activity. It does arise spontaneously in sources that involve watching a film or undertaking a challenge [6], [12], but the activity is not conversation. Various databases derived from TV [13], [14], [15] do show emotion arising from conversations, but because they rarely show both parties, there are important dimensions that they do not capture. Considering face-to-face interaction, there is a similar issue with databases that are unimodal, such as the annotated part of the AIBO database [16] (which is purely audio), and others which are wholly or mainly visual [10], [17]. All of these resources cited do meet other needs: The point is simply that they are not ideal for work on face-to-face, emotionally colored conversations.

A few sources do contain multimodal recordings of both parties in a fluent, emotionally charged conversation, such as the "Green Persuasive" recordings [6] or Canal9 [18]. Arguably, the spontaneous dialog in IEMOCAP may be considered in this category, although the interactions involve acted scenarios (two actors simulate scenes where, e.g., one tells the other she is getting married) [19]. They raise another kind of issue. They suit some research strategies, notably describing human behavior patterns verbally (as psychology has traditionally done) or building components of an affective system (e.g., to recognize user states). However, there are important strategies that they do not support. Specifically, they are not suited to a strategy that tries to progress by building agents that interact as nearly as possible in the way that the recordings show, identifying the problems that arise, and using those to drive progress.

Iterative strategies are commonplace elsewhere as a complement to verbal description and building components. Fluent interaction is a topic where the case for that approach seems particularly clear. Without building systems that interact, or try to, it is all too easy to overlook processes or relationships that are actually crucial for the success of interaction: clearly, the ways components interact, but also rules of timing or responsiveness or coherence that are critical, but not obvious from human data because humans almost never break them. It is difficult to progress with that strategy using material like the Green Persuasive recordings or Canal9 because the interactions depend on competences that are far beyond an artificial agent at present. Specifically, they depend on accurate recognition of fluent speech and subtle interpretation of informal language, all in real time. In contrast to the problems with these verbal competences, the technology does exist to build agents that execute a substantial set of nonverbal skills in real time (as SEMAINE [20] confirmed). That opens the prospect of an iterative strategy, provided that a scenario can be found where the combination of rich nonverbal competences and verbal competences simple enough to be implemented is sufficient to sustain an interaction with a person.

A scenario that seems to meet that requirement was identified some time ago. It is the "Sensitive Artificial Listener," or SAL for short [21]. It is introduced in the next section. The point to be made here is that it offers a way to develop understanding of the nonverbal competences that underpin face-to-face, fluent, emotional interaction by building systems that try to match them. The data described here are set in that scenario, and are designed to let research exploit it.

The choice of scenario also means that the data have some features that are of no great research interest: They are a function of the expedients that give the system its minimal linguistic competence. Hence, research teams who use the data need to ensure that they focus on what is of value, and not on side issues, but that is not a unique problem.

2 The SAL Scenario

The "Sensitive Artificial Listener" scenario had been extensively trialed and refined before SEMAINE adopted it. It was originally suggested by TV chat shows. Not always, but often, hosts use a simple strategy: Invite guests to talk about topics that are emotionally significant for them and encourage (or provoke) them to express the emotion strongly by inserting suitably chosen stock phrases at key points. That model was developed over a substantial period into the scenario considered here.

The interactions involve two parties, a "user" (who is always human) and an "operator" (either a machine or a person simulating a machine). The operator follows (sometimes approximately) a "script" composed of phrases with two key qualities. One is low sensitivity to preceding verbal context: That is, it is usually possible to decide whether a given phrase can be used as the next "move" in a conversation without knowing the words that the user has just said (though it may depend on registering the way they were said). The other is conduciveness: That is, the user is likely to respond to the phrase by continuing the conversation rather than closing it down. Given a repertoire of phrases like that, an operator can conduct a conversation with quite minimal understanding of speech content.

Early experiments with the "script" idea showed that conversation tended to break down unless users felt that the operator had a coherent personality and agenda. Given that the operator's communicative skills center on detecting and expressing emotion, the natural way to define personalities and agendas is in terms of emotions. Hence, we defined subscripts for four "personalities" with appropriately chosen names. Spike is constitutionally angry. He responds empathically when the user expresses anger and critically when he/she expresses any other emotion, which gives the impression that he is "trying" to make the user angry. Similarly, Poppy is happy and "tries" to make the user happy, Obadiah is gloomy and "tries" to make the user gloomy, and Prudence is sensible and "tries" to make the user sensible.

These techniques were evaluated using a system that we have called Powerpoint SAL. The part of the operator was played by a human, who selected appropriate phrases from the prepared script and read them in a tone of voice that suited the character and the context. Its name reflects the fact that the SAL scripts were transcribed onto Powerpoint slides, each one presenting phrases suited to a particular context, accompanied by buttons which allowed the operator to change slides. For instance, if the operator was simulating the Poppy character, and the user's mood was

positive, the operator would navigate to a slide showing phrases that approved and encouraged happiness. He/she would then choose and speak one of them. If the user became angry, clicking a button would bring up a new slide, displaying phrases that Poppy might use to an angry interlocutor. If the user then asked to speak to Spike, another click would bring up a slide showing phrases that Spike might use to an angry interlocutor, and so on.

Recordings made with Powerpoint SAL (in English, Greek, and Hebrew) have been used as data in their own right [22]. What is relevant here is that the work confirmed that users could have quite intense, sustained interactions with an operator whose conversation consisted of phrases from a SAL-type script. It also allowed the scripts to be revised in the light of difficulties. That process generated the scripts used in the program of data collection reported here.

3 Annotation

Annotating emotionally colored conversations is a challenge in its own right. Once again, the techniques described here are part of an iterative process.

Labeling with everyday emotion words faces multiple problems. The states that occur in naturalistic data rarely fit everyday words precisely, it is difficult to capture the rise and fall of emotion, and interrater agreement tends to be low [23]. Labeling with dimensions has obvious attractions, and it forms the core of the scheme used here.

Powerpoint SAL data were annotated using the FEELtrace system [24]. It allows raters to annotate material in terms of two long-established emotion dimensions, valence (how positive or negative the person appears to feel), and activation or arousal (how dynamic or lethargic the person appears to feel) [25]. A rater watches and/or listens to a recording of a target individual, and uses a cursor in an adjacent window to indicate how positive or negative, and active or passive the individual appears to be at any given time. The result is a pair of "traces" which show how perceived valence and activation rise and fall as the recording progresses. Note that for conversation, perceived emotion is what the system needs to know about: It should respond as a person would, even if the person would be wrong [1]. With naturalistic material, the reliability of that approach compares well with verbal ratings [23].

The two-dimensional representation runs throughout Powerpoint SAL. Each character is associated with a region of the space: Obadiah, Spike, and Poppy with different quadrants, Prudence with the center. The same representation is used to organize scripts: The utterances on any given slide are oriented toward a user in one of the same four regions. Hence, a system that can match raters' FEELtrace annotation will be able to match operators' choice of the slide from which the next utterance should be selected. The underlying principle is that annotation should provide the information that a working system needs to make its decisions.

SEMAINE's annotations reflect the same principle. The trace technique was retained, but because there are important distinctions that the dimensions of valence and activation fail to capture, SEMAINE considered a wider set of traces, each using a separate one-dimensional scale [26]. The resulting data provide a basis for assessing the SEMAINE traces in terms of independence, reliability, and, not least, functionality within a working system.

4 Scenarios for SEMAINE Recordings

SEMAINE recordings contrast with earlier SAL material at several levels. Recording quality was much higher (see Section 5.2). Where the operator was a human, it was much easier for the user to regard him/her as a disembodied agent because the two were always in different rooms, communicating via screens, cameras, loudspeakers, and microphones. Most important, the scenario was varied systematically. Three basic scenarios were used: Solid SAL, where human operators play the roles of the SAL characters; Semi-automatic SAL, where a human operator selects phrases from a predefined list but (unlike Powerpoint SAL) the system speaks them; and Automatic SAL, where an automated system chooses sentences and nonverbal signals. These generate a range of interaction types. Solid SAL provides fuller operator-user interaction than Powerpoint SAL, and three variants of Semi-automatic SAL provide progressively less. As a result, the recordings show user responses to different levels of system sophistication.

4.1 Solid SAL

A key objective of the Solid SAL scenario was to record behaviors (mainly nonverbal) that a human operator shows in fluent face-to-face conversation, including their relationships to user behavior—notably backchanneling, eye contact, various synchronies, and so on. That kind of engagement does not occur if the operator is searching a script, or even trying to recover phrases from memory. Hence, the operator in Solid SAL was asked to act in the character of a SAL agent rather than being constrained to using the exact phrases in a SAL script. Acting in character involved adopting the relevant emotional stance (angry for Spike, gloomy for Obadiah, etc.) and, using short, preferably stock utterances with the properties described earlier, low sensitivity to preceding verbal context and conduciveness. The appendix, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/T-AFFC.2011.20, gives sample transcripts that convey the flavor.

Users were encouraged to interact with the characters as spontaneously as possible. There was a single explicit constraint: Users were told that the characters could not answer questions. If they did ask questions, the operator reminded them that the SAL characters could not answer questions. Users talked to the characters in an order of their own choice, and the operator brought the recording session to a close when they had interacted with all four.

The result was not intended to mimic machine human interaction, but it still had important features in common with it. The operator was visible to the participant through a teleprompter screen, and audible through a set of speakers. The indirectness makes it easier to regard the operator as a disembodied agent than it was in Powerpoint SAL. Probably more important, the operator did not behave like a human; he/she followed a simple conversational

agenda, in violation of norms that usually govern human-human interaction.

It is difficult to judge from an abstract description what level of interaction that kind of a scenario might produce. The best indicator comes from the labeling process (described in Section 6), which gave raters several ways of identifying anomalous interactions. Together, they were used in just over 5 percent of ratings, indicating that very little of the user behavior was either contrived or disengaged.

Twenty-four recording sessions used the Solid SAL scenario. Recordings were made of both the user and the operator and there were usually four character interactions in each recording session, providing a total of 95 character interactions and 190 video clips.

4.2 Semi-Automatic SAL

Semi-automatic SAL was similar to Powerpoint SAL in that a human operator chose phrases from a predefined script. These were made available to her/him through a Graphical User Interface based on the powerpoint SAL model. Navigation buttons allowed him/her to bring up a page of utterances related to the current character and the user's current emotional state. When he/she clicked on a phrase, it was then played using a prerecorded audio file spoken by an actor whose voice had been judged appropriate for the character. As before, the user heard through loudspeakers and looked at a teleprompter. Its screen showed a simplified face designed to keep users looking in the general direction of the camera behind it. In order to hold attention, the spectrum of the speech was placed just below the "mouth" of the face. The fact that it changed in time with the speech helped to create the impression that the speech was associated with it.

The Semi-automatic SAL scenario included three variants which gave the operator progressively less feedback from the user. In the baseline condition (Experiment 1), the operator both saw and heard the user, and could therefore use information from both the user's words and his/her nonverbal signals to choose an appropriate utterance. In the remaining variants, the operator had to choose utterances on the basis of video with an audio either switched off (Experiment 2) or with an audio filtered to remove verbal information (Experiment 3). The filter cut out frequencies between 350 and 4,000 Hz, which leaves prosody largely intact, but only occasional words can be made out. The degradation made it harder for the operator to avoid inappropriate choices of the kind that the system used in the Automatic SAL scenario would necessarily make (because it does not use linguistic information), and resulted in recordings where users showed various signs of communication breakdown.

In the first experiment, 11 Semi-automatic SAL recordings provided 44 character sessions with a procedure directly comparable to Solid SAL. Experiments 2 and 3 used degraded versions of Semi-automatic SAL in which two of the four character sessions were with the full Semi-automatic SAL system while the other two were degraded. A further 25 sessions took place (13 in Experiment 2 and 12 in Experiment 3) with differing degrees of degradation of information to the operator. The operator videos consist only of the operator interacting with the interface and show little of interest regarding the conversational interaction; therefore only the user videos are included database. There were four character sessions for each recording in the Semi-automatic SAL experiments; these add a further 144 videos to the database (see Table 2 for an overview).

Fig. 1. The four SAL character avatars. Clockwise from top-left: Spike, Poppy, Prudence, and Obadiah.

4.3 Automatic SAL

In the fully automatic SAL recordings, the utterances and nonverbal actions executed by the SAL Character were decided entirely automatically by the current version of the SEMAINE project system. The system is described in detail elsewhere [20], but a brief overview is given here for completeness.

The user sat in front of a teleprompter, as in Semi-automatic SAL. The system's sensors were a grayscale camera and microphones (see Section 5.2 for details). From the video input, the system detected when a person's face is present, significant gestures (head nods and shakes), and facial actions (smiles, eyebrow raising and lowering, mouth opening). From the audio, it detected the presence or absence of user speech, emotion-related prosodic features, and words that could be recognized with high confidence. These fed into different channels, some governing actions (e.g., initiating an utterance when the user stopped speaking, or nodding in response to the user's nod) and some using the information to infer the user's emotions (using the dimensional descriptors described in Section 6.1.1). Potentially relevant utterances were chosen on the basis of conversational norms (e.g., avoiding repetition), key words if they were available, and inferred emotion.

The system's outputs were audiovisual. Visual output consisted of avatars designed to represent the SAL characters (see Fig. 1), with movements and expressions controlled by the analyses described above. Audio output consisted of phrases from the relevant script spoken by a synthetic unit selection voice, with a different voice for each character. Behavior depended critically on parameters governing weightings of different information sources and rules,

Fig. 2. Images of the recording setup for both the User (left) and the Operator (right) Rooms.

response magnitudes and latencies, and so on. These were adjusted during testing, and account for the main differences between the versions used in data collection (see below).

Participants in the experiments interacted with two versions of the system, one with the best set of nonverbal skills available and one with a degraded set (hence, they interacted with each of the four characters twice). Sessions were limited to approximately 3 minutes; or if the participant did not engage with the system, they were ended after a minimum of 1.5 minutes. There were three iterations of this procedure using five versions of the system, two degraded versions that removed affective cues and three iterations of the fully operational version of the SAL system, based on variations of SEMAINE system 3.0.1. An initial experiment examined the effect of the system's perceptual abilities, comparing a full version of the system based with a degraded version which ignored the user's actual emotional state, and chose its responses at random. Fifteen participants were tested using this configuration adding 120 character sessions to the database. A second experiment used two different system versions; a new full version (with some bug fixes) and a new degraded system that removed most of the system's affective output, no backchanneling, or facial emotional information and random utterance selection, and flat affect in the agent voices. This examined the utility of the emotional information in the characters. This added 240 character sessions to the database. The third experiment used the same degraded system as experiment 2, and a different full version with an improved dialogue management system and further bug fixes. This added a further 240 character sessions to the database. Additionally, five pilot sessions (recorded between experiments 2 and 3) added a further 40 videos to the database. A screen grab of the video output of the Agent computer added 324 agent videos to the database. In total the Automatic SAL scenario provides 964 videos to the database. Examples of Automatic SAL character interactions are available online [27].

5 Participants and Procedure

5.1 Participants and Procedure

The data set features 150 participants. The youngest participant was 22, the oldest 60, and the average age is 32.8 years old (std. 11.9). Thirty-eight percent are male. Participants come from eight different countries: Most were from a Caucasian background. Participants were undergraduate and postgraduate students. The overwhelming majority took part in only one scenario. Before taking part, participants were briefed about the project and provided written consent for use of the recordings. Typical session duration for Solid SAL and Semi-automatic SAL was about 30 minutes, with an approximate interaction time of 5 minutes per character, though there were considerable individual variations. Participants were told to ask for a different character when they got bored, annoyed, or felt they had nothing more to say to the character. The operator could also suggest a change of character if an interaction was unusually long or had reached a natural conclusion. The Automatic SAL session duration was about 1 hour, with eight character interactions of approximately 3 minutes each. The participants interacted with two versions of the system with an intervening 10-15 minute period in which they completed psychometric measures.

The interaction procedure was the same throughout the experiments. Participants entered the recording studio, where they sat in the user room and put on their head microphone. The operator took her/his place in a separate recording room and recording starts, as in Fig. 2 (details of how face-to-face conversations were maintained while recordings were made are given in the following section). The operator/agent recited a brief introduction script and the interaction began.

After each session, there was a debriefing session, allowing the user to ask more about the system.

5.2 Synchronized Multisensor Recording Setup

The database was created with two distinct types of use 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. Second, the data are intended to be used for the creation of machines that can interact with humans by learning how to recognize social signals. These considerations guided the decisions on the choice of sensors, and how the sensors are placed.

Sensors. Video was recorded at 49.979 frames per second and at a spatial resolution of 780×580 pixels using AVT Stingray cameras. Both User and Operator were recorded from the front by both a grayscale camera and a color camera. In addition, the User was recorded by a grayscale camera positioned on one side of the User to capture a profile view. An example of the output of all five cameras is shown in Fig. 3. The reason for using both a color and a grayscale camera is directly related to the two target audiences. A color camera sacrifices spatial resolution for color. Machine vision methods usually prefer the sharper grayscale image over a blurrier color image. For humans, however, it is more informative to use the color image [28].

Fig. 3. Frames grabbed at a single moment in time from all five video streams. The Operator (left) has HumanID 7 and the User (right) has HumanID 14. Shown is the 3,214th frame of the 19th recording.

To record User and Operator speech, there were two microphones per person: one placed on a table in front of the User/Operator and the second worn on the head by the User/Operator. The wearable microphones were AKG HC-577-L condenser microphones, while the room microphones were AKG C1000-S microphones. This results in four audio channels. The wearable microphone was the main source for capturing the speech and other vocalizations made by the User/Operator, while the room microphones were used to model the background noise. Audio was recorded at 48 kHz and 24 bits per sample.

Environment. The User and the Operator were located in separate rooms. They heard each other through speakers, which played the audio recorded by the wearable microphone of their conversational partner. They saw each other through teleprompters. Each teleprompter contained two cameras recording a person's frontal view placed behind the semireflecting mirror. That allowed the User and the Operator to have the sense of looking each other in the eye. A pilot test used cameras placed on top of a screen which showed the other party's face, but that did not give an impression of eye contact, and greatly reduced the sense of a direct communication. Professional lighting was used to ensure an even illumination of the faces. Images of the two rooms can be seen in Fig. 2.

Synchronization. To do a multisensory fusion analysis of the recordings, it is essential that all sensor data are recorded with the maximum synchronization possible. A system developed by Lichtenauer et al. [29] was used to achieve that. It uses the trigger of a single camera to accurately control when all cameras capture a frame. This ensures 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 synchronized audio and video sensor data with a maximum time difference between data samples of 25 μsec.

Data compression. The amount of raw data generated by the visual sensors is large: 959 character interactions, lasting 5 minutes on average, recorded at 49.979 frames/second at a temporal resolution of $780 * 580$ pixels with 8 bits per pixel for five cameras, would result in 29.6 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 were compressed using the (lossy) H.264 codec and stored in an avi container. The video was compressed to 440 kbit/s for the grayscale video and to 500 kbit/s for the color video. The recorded audio was stored without compression, because the total size of the audio signal was much smaller.

TABLE 1
Solid SAL Recordings

Users	Sessions/User	Total Time	Annotators
Solid SAL User	95/24	475	6+
Solid SAL Operator	95/4	475	1

5.3 Summary of the SEMAINE Recordings

Tables 1, 2, and 3 summarize the recordings.

6 Annotation and Associated Information

6.1 Trace Annotation of Participant States

Building on experience with Powerpoint SAL, trace-style continuous ratings were used to record raters' impressions of user states—primarily emotion-related—that appeared potentially relevant to controlling an automatic system. The specific traces were chosen in consultation with the SEMAINE members involved in building automatic SAL. The main tracing system (applied to Solid SAL and Semiautomatic SAL recordings) involved two stages. Five core traces (described in Section 6.1.1 below) were provided by every rater for every clip. After making those core traces, raters were offered a menu of optional descriptors (listed in

TABLE 2
Semi-Automatic SAL Recordings

Experiment & System	Sessions/User	Approximate Total Time	Annotators
Experiment 1			
Full audio	44/11	220	2
Experiment 2			
Full audio	26/13	130	1
No Audio	26/13	130	1
Experiment 3			
Full audio	24/12	120	1
Degraded Audio	12/12	60	1
Degraded Audio & No Vision	12/12	60	1

Time is measured in minutes.

TABLE 3
Automatic SAL Recordings

Experiment & System	Sessions/User	Approximate Total Time	Annotators
Experiment 1			
Full 1	60/15	180	1
Degraded 1	60/15	180	1
Experiment 2			
Full 2	120/30	360	1
Degraded 2	120/30	360	1
Experiment 3			
Full 3	120/30	360	1
Degraded 2	120/30	360	1
Pilots	40/5	120	1
Agent video	324/81	972	1

Time is measured in minutes.

Section 6.1.2). From it, each rater independently chose four that he/she felt were definitely exemplified in the clip. More than four could be chosen if there seemed to be strong instances of more than four categories, but that rarely happened. The rater then made a new trace for each of his/her choices, indicating how strongly the user exhibited the state in question from moment to moment. Hence, each rater provided nine traces in all—five core and four optional—for each clip.

6.1.1 Core Dimensions

The five core dimensions were valence, activation, power, anticipation/expectation, and intensity. The first four reflect an influential recent study [30] which argues that they account for most of the distinctions between everyday emotion categories. The first two have already been introduced. The power dimension subsumes two related concepts, power and control. These are not the same conceptually—power is mainly about internal resources, control is about the relationship between those resources and external factors. In practice, raters find it natural to make a composite judgment, dealing with the balance between the two. Anticipation/Expectation also subsumes various concepts that can be separated—expecting, anticipating, being taken unawares. Again, people find it intuitively meaningful to make a composite judgment, related to control in the domain of information. The last dimension, overall intensity, is about how far the person is from a state of pure, cool rationality, whatever the direction. Logically, one might hope that it could be derived from the others, but that is something to be tested rather than assumed. This trace serves a function that is handled differently in other databases. Periods when the person is judged to be unemotional are marked by low values in the intensity trace.

6.1.2 Optional Descriptors

The optional traces dealt with categories from everyday language or psychological theory, and were identified by SEMAINE partners as potentially relevant to system decisions. They were of four main types.

Basic emotions. Seven labels of this type were offered: fear, anger, happiness, sadness, disgust, contempt, and amusement. It is important to know whether they can be clearly identified in this kind of material or derived from other descriptors because they are integral to existing techniques for, e.g., generating facial expressions. Most of the items from the best-known list of basic emotions, Ekman's, were included as options. Surprise was excluded because tracing it would almost inevitably duplicate information that was already in the expectation/anticipation trace, at the cost of information about another category. Conversely, amusement is clearly an important category in this kind of conversation. This is the most convenient place to include it (and some authors do consider it a basic emotion, e.g., [31]).

Epistemic states. These states were highlighted by Baron-Cohen et al. [32], and have been viewed within the machine perception community as a significant resource for describing everyday emotion [5]. They are relatively self-explanatory. The options of this type were:

- certain/not certain,
- agreeing/not agreeing,
- interested/not interested,
- at ease/not at ease,
- thoughtful/not thoughtful,
- concentrating/not concentrating.

Interaction process analysis. The descriptors offered here are a subset of the system of categories used in Interaction Process Analysis [33]. IPA categories are used in dialogue management, and so ability to recognize instances would be practically useful. The labels offered were five pairs:

- Shows Solidarity, Shows Antagonism,
- Shows Tension, Releases Tension,
- Makes Suggestion, Asks for Suggestion,
- Gives Opinion, Asks for Opinion,
- Gives Information, Asks for Information.

Validity. The final set of labels was intended to highlight cases where the user was not communicating his or her feelings in a straightforward way. Among other things, that affects the way the material should be used carefully or not at all in a training context. The labels offered were:

- *Breakdown of engagement.* This seeks to identify periods where one or more participants are not engaging with the interaction. For example, they are thinking of other things, looking elsewhere, ignoring what the other party says, rejecting the fiction that they are speaking to, or as SAL characters rather than to or as the actual people involved.
- *Anomalous simulation.* This label 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. The main hallmark is that the expressive elements do not go together in a fluent or coherent way—they are protracted or separated or incongruous.
- *Marked sociable concealment.* This is concerned with periods when it seems that a person is feeling a definite emotion, but is making an effort not to show it. In contrast to the two categories above, this is something that occurs in everyday interaction. It is an aspect of what Ekman et al. [34] call display rules.
- *Marked sociable simulation.* This 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. Again, this is something that occurs in everyday interaction. People simulate interest or friendliness or even anger that they do not feel, not necessarily to deceive, but to facilitate interaction.

6.1.3 Traces of Engagement

The scheme described so far was applied to Solid SAL and Semiautomatic SAL recordings. Time prevented applying it to automatic SAL recordings. However, a related procedure provided information that is relevant both to system evaluation and to wider research questions. As automatic SAL interactions took place, a rater watching a live video feed of the interaction traced the user's apparent engagement in the interaction.

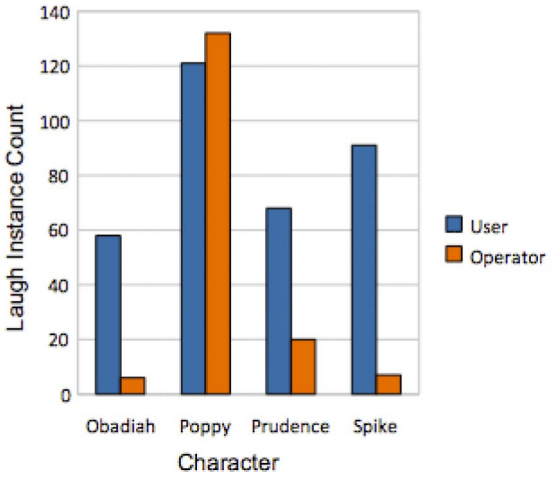


Fig. 4. Instances of user and operator laughter for each character in Solid SAL recordings 1-21.

6.1.4 Amount of Annotation

The amount of annotation provided reflects the time available. Solid SAL was completed first and has the largest body of annotation followed by Semi-automatic SAL. Both are annotated with the five core dimensions and four optional categories. Automatic SAL has the least annotation, with traces of engagement only. The number of traces are as follows:

Solid SAL. For user clips, four sessions have been annotated by eight raters, 17 by 6, and the remainder by at least three. For operator clips three have been annotated by four raters, the rest by one rater.

Semi-automatic SAL. Eleven user sessions have been annotated by two raters.

Automatic SAL. All sessions have been annotated by a single rater.

Annotation is being extended gradually.

6.2 Transcripts

Of the 24 Solid SAL sessions 21 were fully transcribed creating 75 transcribed character interactions. The transcripts were time aligned with detected turn taking changes. None of the user interactions in the Semiautomatic SAL or Automatic SAL sessions have been transcribed but the operator utterances are automatically recorded and made available as log files.

6.3 Laughs

An initial subset of laughter was identified in the transcription process. This was added using the SEMAINE laugh detector which was manually corrected and aligned. These laughs are included in the aligned transcripts with the time of occurrence and the annotation <LAUGH>. User laughter was present in 56 out of 66 transcribed character interactions. The rates of laughter varied by character, and number of instances of laughter for each character, for both user and operator, can be seen in Fig. 4.

6.4 Nods and Shakes

Instances of nods and shakes were specifically identified within the database. One hundred fifty-four nods and 104 head shakes were annotated by two raters, using two annotation strategies. The first was a subset of the main SEMAINE annotations deemed most appropriate to nods and shakes (valence, arousal, agreeing/disagreeing, at ease/not at ease, solidarity, antagonism, understanding). The second used annotations derived from McClave [35]; these were Inclusivity, Intensification, Uncertainty, Direct quotes, Expression of mental images of characters, Deixis and referential use of space, Lists or alternatives, Lexical repairs, Backchanneling requests. The results of preliminary analysis and greater detail regarding the annotations can be found in [36].

6.5 FACS Annotation

FACS is a coding scheme developed to objectively describe facial expressions in terms of visible muscle contractions/relaxations. To be able to test existing and/or new automatic FACS coding systems, eight character interactions received a sparse FACS coding [37]. Instances were labeled for the presence of Action Units, specified by frame number and whether they occur in combination with other Action Units or in isolation. Three certified FACs coders at QUB annotated selected frames in the eight interactions, obtaining 577 facial muscle action (Action Unit) codings in 181 frames, which was deemed to be sufficient to perform preliminary tests on this database. Action Unit annotations are available with the database.

7 Evaluation and Analysis of the Data

7.1 Quality of Interaction

One of the key evaluation issues is the quality of the interaction shown: whether it is natural, representative of foreseeable types of human-machine interaction, or simply contrived. SEMAINE incorporated various ways to answer those questions.

Impressionistic judgments cannot be ignored. The verbal content of the exchanges gives some indication. The transcripts in the Appendix, available in the online supplemental material, illustrate what happened in the scenarios involving the most and least human-like operators, Solid SAL, and Automatic SAL. In Solid SAL, the operator is single-minded, but it is clear that that there is a lively interchange. In Automatic SAL, the impression is that it is hard to read the operator's train of thought, but a cooperative user (like this one) can find plausible directions to follow.

The transcripts obscure the nonverbal behaviors which signal participants' engagement or lack of it. In Solid SAL, engagement was overwhelmingly the norm. The labeling process incorporated several ways of identifying anomalous interactions. Together, they were used in just over 5 percent of ratings, indicating that very little of the user behavior is either contrived or is engaged.

In Semi-automatic and Automatic SAL sessions, it was clear that interaction sometimes broke down. Whether that is a problem depends on the frequency of breakdown. Several techniques were used to identify sessions where problems arose. The experimental procedure in Semi-automatic and Automatic SAL included three questions to users about the quality of the interaction: "How naturally do you feel the conversation flowed?" "Did you feel the Avatar said things completely out of place? If yes how often?" and "How much did you feel you were involved in

TABLE 4
Alpha Coefficient for Functionals Associated with Each Trace Dimension (* Indicates Alpha >0.6—the Lowest Value Commonly Considered Acceptable, ** Indicates Alpha >0.7—Almost Always Considered Acceptable, † Indicates Nonacceptable Values)

	Intensity	Valence	Activ	Power	Expect
Mean all	0.74 **	0.92 **	0.73 **	0.68 *	0.71 **
sd bins	0.83 **	0.75 **	0.65 *	0.61 *	0.68 *
min bin	0.23 †	0.90 **	0.43 †	0.43 †	0.43 †
median bin	0.72 **	0.91 **	0.72 **	0.67 *	0.68 *
max bin	0.74 **	0.92 **	0.73 **	0.68 *	0.71 **
AveMagnRise	0.74 **	0.49 †	0.53 †	0.39 †	0.58 †
SDMagnRise	0.74 **	0.60 *	0.63 *	0.32 †	0.59 †
MaxMagnRise	0.75 **	0.56 †	0.64 *	0.25 †	0.63 *
AveMagnFall	0.68 *	0.45 †	0.55 †	0.55 †	0.51 †
SDMagnFall	0.66 *	0.45 †	0.63 *	0.60 *	0.49 †
MinMagnFall	0.60 *	0.46 †	0.59 †	0.60 *	0.41 †

TABLE 5
Distribution of Optional Traces for the 13 Most Used Options (No Others Reach 5 per Character or 10 across Characters)

Optional Trace	Obadiah	Poppy	Prudence	Spike
Gives Information	10	20	19	9
Agreeing	15	11	15	15
Amusement	8	14	13	12
Gives Opinion	12	7	9	11
Thoughtful	10	9	8	4
At Ease	5	6	7	9
Certain	4	5	9	4
Happiness	2	15	5	1
Sadness	13	1	1	0
Anger	1	0	2	8
Shows Antagonism	0	1	1	6
Contempt	0	0	1	5
Interested	3	3	2	2

the conversation?" The sessions also included a "Yuck button," which users were asked to press when the interaction felt unnatural or awkward. In both Semi-automatic and Automatic SAL, each interaction was followed by an open ended invitation to state the way the user felt about the conversation. In Automatic SAL, an additional layer was available where an observer used a FEELtrace-type scale to rate each participant's apparent level of engagement.

The database includes information from all these sources. A useful overall indicator is that in the final session with automatic SAL, average self-ratings of engagement with Poppy, Spike, Obadiah, and Prudence were, respectively, 6.2, 6.4, 7. 1, and 6.1 on scale from 0 (none) to 10 (complete). Hence, from the users point of view, a substantial proportion of the interactions were thoroughly engaging. In contrast, the malinteractions provide data relevant to recognizing problems that are likely to be important in human-machine interaction for the foreseeable future.

7.2 Reliability of Main Traces

The trace set available for Solid SAL allowed reliability to be measured in two stages. The first considered relationships between clips, using functionals derived automatically from each trace of each clip (mean, standard deviation, average magnitude of continuous rises, etc.). Correlations can then be used to measure agreement between the list of (for example) mean valence ratings, one for each clip, produced by any one rater, and the corresponding list from any other. From that, the standard Cronbach's alpha measure of agreement can be calculated. Table 4 summarizes the results. Overall, the findings confirm that most of the ratings are reliable, though not necessarily in the same respects. Average and maximum level are rated reliably for all the traces except power, and there the effect is just short of the standard level. Beyond that, judgments of intensity and valence seem to show consistent patterns of rises, though in different respects. For intensity, it is the magnitude of the rises on which raters agree. For valence, it is their frequency.

It is more difficult to measure an intraclip agreement (that is, agreement between raters on the way a single measure, say valence, rises and falls in the course of a single clip). As a straightforward option, we reduced each trace to a list of values (averages over 3 sec bins), and calculated the correlations between all the resulting pairs of lists. Again, Cronbach's alpha coefficients can be derived from the correlations.

Alpha was calculated for 305 sets of traces, each describing a single clip on one of the core dimensions. Less than 10 percent fail to reach the standard criterion of $\mathrm{alpha} = 0.7$, and more than 70 percent meet a stringent criterion of alpha > 0.85. There are reasons to be wary of alpha as a measure with this kind of data, and for that reason we developed an alternative method, called QA, for Qualitative Agreement. The relevant point here is just that although the basis for calculating agreement is completely different, and specifically avoids the problem assumptions, it gives very similar conclusions about the overall level of agreement in the sample. Details of the method and the results are in [38]. There are some differences between the different types of trace. For intensity, valence, and power, over 60 percent of trace sets meet the stringent criterion. The figure is much lower for activation (51 percent), and much higher for expectation (86 percent). These differences invite exploration. Again, Cowie and McKeown [38] give more detail.

7.3 Distribution of Optional Traces

The "optional" trace categories indicate where raters felt that particular qualitative descriptors applied, and show how the chosen states appeared to change over time. Table 5 provides an overview of the most used options for each of the characters (for the sake of balance, only data from the six raters who traced all the clips are included). Responses are considered for each character because the different characters do get quite different responses—for instance, sadness is rare overall, but quite common in interaction with Obadiah, and showing antagonism is rare overall, but common with Spike.

It is clear that the vast majority of responses describe a few core positions relative to the exchange. After those come emotions directly related to the character of the operator. Very few of the other categories feature at all often. The implication is that most of the information that tracing can provide can be captured by quite a modest number of traces. Considering intercorrelations among

traces may show that it can be reduced further. That is a research question that the data can be used to explore.

8 Automatic Analysis of the Database

The quality and scale of the SEMAINE corpus provides an opportunity to develop new ways of automatically analyzing human behavior by detecting social signals. The synchronous high quality audio and video streams, combined with the large amount of manual annotations, allow audio and computer vision researchers to develop new systems and evaluate them on naturalistic data. It has already been used in that capacity for a number of other related projects, and their results illustrate the potential.

Jiang et al. [39] reported on facial muscle action (FACS Action Units, AUs) detection on the SEMAINE data. They compared two appearance descriptors (Local Binary Patterns and Local Phase Quantization), and found that between the two Local Phase Quantization performed best. They were able to detect 7 AUs with an average F1-measure of 76.5 percent. However, this was tested on only eight sessions of only two subjects. The authors found that there was a big difference in performance between the two subjects. They reported that the temporal extension of LPQ, called LPQ-TOP, attained the highest performance.

Gunes and Pantic [40] proposed a system to automatically detect head nods and shakes, and continued to detect the affective dimensions arousal, expectation, intensity, power, and valence. To detect the head actions nodding and shaking, they first extracted global head motion based on optical flow. The detected head actions together with the global head motion vectors were then used to predict the values of the five dimensions labeled in all recordings (arousal, expectation, intensity, power, and valence). In the process they addressed the notoriously difficult problem of differences in interpretation by different observers [41] by modeling each annotator directly, independent of the others.

Nicolaou and Pantic [42] developed a method to use the continuous dimensional labels of multiple annotators to automatically segment videos. Their aim was to develop algorithms that produce ground-truth by maximizing intercoder agreement, identify transitions between emotional states, and that automatically segment audiovisual data so it can be used by machine learning techniques that require presegmented sequences. They tested their approach on the SEMAINE corpus and reported that the segmentation process appeared to be effective, with the segments identified by their algorithm capturing the targeted emotional transitions well.

Eyben et al. [43] used the SEMAINE corpus to first detect a range of nonverbal audio-visual events and then use these to predict the values of five dimensions: Valence, Arousal, Expectation, Intensity, and Power. The visual events they detected were face presence, facial muscle actions (FACS Action Units), and the head actions nodding, shaking, and head tilts. The acoustic events they detected were laughter and sighs. The events were detected on the basis of a short temporal window, and combined into a single bag-of-words feature vector. They reported that results using this string-based approach were at least as good as the traditional signal-based approaches, and performed best for the dimensions Valence and Expectation. They also reported that the detection of events always adds information relevant to the problem, that is, when the detected events are combined with the signal-level features the performance always increases.

9 Availability

The SEMAINE data set is made freely available to the research community. It is available through a web-accessible interface with url http://semaine-db.eu/.

9.1 Organization

Within the database, the data are organized in units that we call a *Session*, in which the User speaks with a single Character. There are also two special sessions per recording, the *recording_start* and *recording_end* sessions, where the User/Operator prepares to do the experiment or ends it, and in Semi-automatic and Automatic SAL there are evaluation recordings. 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 or detectors which sense the presence of a User.

The number of sensors associated with each session depends on the originating scenario: Solid SAL recordings have nine sensors associated with them, while all other scenarios have seven. We call the sensor database entries *Tracks*. Nine of these are the five camera recordings and the four microphone recordings (see Section 5.2). In addition, each session has two lower quality audio-visual Tracks, showing the frontal color recordings of the User and the Operator, respectively. Both have audio from both speakers. The fact that these have both audio and video information makes them useful for 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. A standard *balance* slider allows a rater to choose who to listen to. The low-quality tracks are also small for convenient 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. A screenshot of the web interface can be seen in Fig. 5. 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 people 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.

Approximately one-third of the recorded data are being withheld from public access to allow for benchmarks procedures to be set up and for the organization of

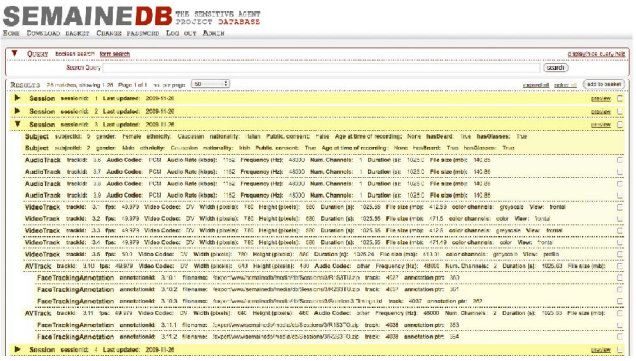


Fig. 5. Data organization of the database.

challenges similar to the Interspeech audio-analysis series (e.g., [44]) and the FERA facial expression recognition challenge [45]. The database also defines a partitioning of the publicly available data into a training, development, and test set. The training set would be used by researchers to train their systems with all relevant parameters set to a specific value, while the development set would then be used to evaluate the performance of the system given these parameters. The partitioning information is specified in two text files available from the website.

9.2 Search

To give researchers ready access, 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 have become very familiar with the database. 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 more. Once a search is concluded, the user can inspect the properties of the returned sessions, tracks, and annotations, and/or watch a preview of all the returned video tracks. A screenshot of the search interface can be seen in Fig. 6.

10 Conclusion

The SEMAINE database is a point of departure for several distinct kinds of development.

Most directly, it provides a resource that computational research can use immediately. Section 8 has indicated that is already under way. A new avenue is opened by information about the user's level of engagement in both Automatic and Semiautomatic SAL. Recognizing level of engagement is a natural challenge and probably not too intractable.

Beyond that, it is natural to add new types of labeling to the recordings. That has various levels. The kind of tracing that has been applied to Solid SAL should be extended to Automatic and Semi-automatic SAL recordings. More radically, fuller annotation of gestures in the recordings would open the way to a range of analyses. Multiple types of gesture are present—facial movements, head nods and shakes, and laughs. The quality of the material means that identification could be automated to a large extent, providing what would be by contemporary standards a very large source of information on the contingencies between these various elements, and their relationship to the parties' emotions and engagement.

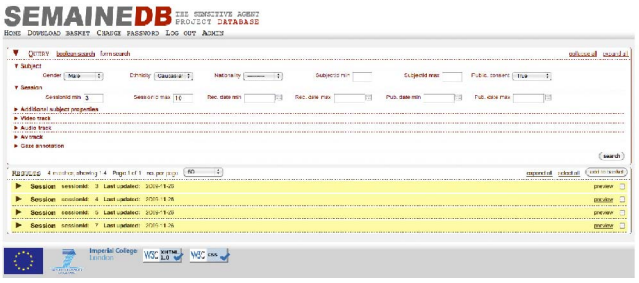


Fig. 6. Form search: some options and the search results.

These developments are of interest to the human sciences as well as to computing. For example, substantial theoretical issues hinge on the way facial gestures appear in spontaneous emotional expression, but the scarcity of naturalistic material and the labor of identifying facial actions has made it difficult to draw strong conclusions [46], [47]. The issue affects not only the generation of emotion-related signals, but also the mechanisms needed to recover information from such signal configurations [48]. SAL data offer a realistic prospect of addressing these questions.

Deeper questions hinge on the point, emphasized throughout, that interacting with an artificial agent is not the same as interacting with a human. Up to a point, they can be treated as separate problems. However, the contrast also offers new ways to expose a multitude of factors that make human-human interaction what it is, but whose effect is usually so automatic that we do not realize they are there.

Last but not least, the SEMAINE approach to data collection provides a model that it makes sense to generalize. If, as seems likely, the expression of emotion is highly context-specific, then there is little alternative to careful iterative construction of databases, working through simulations to full prototype systems. It would be easier if one could move directly from databases showing general examples of emotion to systems that carried out specific functions, but in this area, nature seems to have elected not to make life easy.

Acknowledgments

This work has been funded by the European Community's 7th Framework Programme [FP7/2007-2013] under grant agreement no. 211486 (SEMAINE).

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine,* vol. 18, no. 1, pp. 32-80, Jan. 2001.

[2] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond Emotion Archetypes: Databases for Emotion Modelling Using Neural Networks," *Neural Networks,* vol. 18, no. 4, pp. 371-88, 2005.

[3] R. Cowie and M. Schröder, "Piecing Together the Emotion Jigsaw," *Machine Learning for Multimodal Interaction,* vol. 3361, pp. 305-317, 2005.

[4] R. Cowie, E. Douglas-Cowie, J.-C. Martin, and L. Devillers, "The Essential Role of Human Databases for Learning in and Validation of Affectively Competent Agents," *A Blueprint for Affective Computing: A Sourcebook and Manual,* K.R. Scherer, T. Bänziger, and E. Roesch, eds., pp. 151-165, Oxford Univ. Press, 2010.

[5] S. Afzal and P. Robinson, "Natural Affect Data—Collection & Annotation in a Learning Context," *Proc. Third Int'l Conf. Affective Computing and Intelligent Interaction and Workshops,* pp. 1-7, 2009.

[6] R. Cowie, E. Douglas-Cowie, I. Sneddon, M. McRorie, J. Hanratty, E. McMahon, and G. McKeown, "Induction Techniques Developed to Illuminate Relationships Between Signs of Emotion and Their Context, Physical and Social," *A Blueprint for Affective Computing: A Sourcebook and Manual,* K.R. Scherer, T. Bänziger, and E. Roesch, eds., pp. 295 -307, Oxford Univ. Press, 2010.

[7] D. Heylen, D. Reidsma, and R. Ordelman, "Annotating State of Mind in Meeting Data," *Proc. Workshop Programme Corpora for Research on Emotion and Affect,* pp. 84-170, 2006.

[8] P. Ekman and W.V. Friesen, *Pictures of Facial Affect.* Consulting Psychologists Press, 1976.

[9] M. Kienast and W. Sendlmeier, "Acoustical Analysis of Spectral and Temporal Changes in Emotional Speech," *Proc. ISCA Tutorial and Research Workshop Speech and Emotion,* 2000.

[10] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis," *Proc. IEEE Fourth Int'l Conf. Automatic Face and Gesture Recognition,* pp. 46-53, 2000.

[11] T. Bänziger and K.R. Scherer, "Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The Gemep Corpus," *Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction,* pp. 476-487, 2007.

[12] J. Hanratty, "Individual and Situational Differences in Emotional Expression," PhD dissertation, School of Psychology, Queen's Univ. Belfast, 2010.

[13] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A New Emotion Database: Considerations, Sources and Scope," *Proc. ISCA Tutorial and Research Workshop Speech and Emotion,* 2000.

[14] S. Abrilian, L. Devillers, S. Buisine, and J.-C. Martin, "Emotv1: Annotation of Real-Life Emotions for the Specification of Multimodal Affective Interfaces," *Proc. 11th Int'l Conf. Human-Computer Interaction,* July 2005.

[15] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 865-868, 2008.

[16] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong, "'You Stupid Tin Box'—Children Interacting with the Aibo Robot: A Cross-Linguistic Emotional Speech Corpus," *Proc. Fourth Int'l Conf. Language Resources and Evaluation,* pp. 171-174, 2004.

[17] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-Based Database for Facial Expression Analysis," *Proc. IEEE Int'l Conf. Multimedia and Expo,* p. 5, 2005.

[18] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A Database of Political Debates for Analysis of Social Interactions," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction,* pp. 1-4, 2009.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *J. Language Resources and Evaluation,* vol. 42, no. 4, pp. 335-359, 2008.

[20] M. Schröder et al., "Building Autonomous Sensitive Artificial Listeners," *IEEE Trans. Affective Computing,* under revision.

[21] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, "The Sensitive Artificial Listener: An Induction Technique for Generating Emotionally Coloured Conversation," *Proc. Workshop Corpora for Research on Emotion and Affect,* 2008.

[22] G. Caridakis, K. Karpouzis, M. Wallace, L. Kessous, and N. Amir, "Multimodal User's Affective State Analysis in Naturalistic Interaction," *J. Multimodal User Interfaces,* vol. 3, pp. 49-66, 2010.

[23] R. Cowie and R. Cornelius, "Describing the Emotional States that Are Expressed in Speech," *Speech Comm.,* vol. 40, nos. 1/2, pp. 5-32, 2003.

[24] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder, "'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time," *Proc. ISCA Tutorial and Research Workshop Speech and Emotion,* 2000.

[25] J. Russell and L. Barrett, "Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant," *J. Personality and Social Psychology,* vol. 76, no. 5, pp. 805-819, 1999.

[26] R. Cowie, C. Cox, J.-C. Martin, A. Batliner, D. Heylen, and K. Karpouzis, "Issues in Data Labelling," *Emotion-Oriented Systems: The Humaine Handbook,* R. Cowie, C. Pelachaud, and P. Petta, eds., pp. 213-241, Springer-Verlag, 2011.

[27] G. McKeown, "Chatting with a Virtual Agent: The Semaine Project Character Spike [Video File]," http://youtu.be/6KZc6e_EuCg, 2011.

[28] P. Valdez and A. Mehrabian, "Effects of Color on Emotions," *J. Experimental Psychology-General,* vol. 123, pp. 394-408, 1994.

[29] J. Lichtenauer, J. Shen, M. Valstar, and M. Pantic, "Cost-Effective Solution to Synchronised Audio-Visual Data Capture Using Multiple Sensors," *Proc. IEEE Int'l Conf Advanced Video and Signal Based Surveillance,* pp. 324-329, 2010.

[30] J. Fontaine, S.K.R., E. Roesch, and P. Ellsworth, "The World of Emotions Is Not Two-Dimensional," *Psychological Science,* vol. 18, no. 2, pp. 1050-1057, 2007.

[31] P. Ekman, "Basic Emotions," *Handbook of Cognition and Emotion,* pp. 45-60, John Wiley, 1999.

[32] S. Baron-Cohen, O. Golan, S. Wheelwright, and J.J. Hill, *Mind Reading: The Interactive Guide to Emotions.* Jessica Kingsley Publishers, 2004.

[33] R.F. Bales, *Interaction Process Analysis: A Method for the Study of Groups.* Addison Wesley, 1951.

[34] P. Ekman, W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, and P.E. Ricci-Bitti, "Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion," *J. Personality and Social Psychology,* vol. 53, no. 4, pp. 712-717, 1987.

[35] E. McClave, "Linguistic Functions of Head Movements in the Context of Speech," *J. Pragmatics,* vol. 32, no. 7, pp. 855-878, 2000.

[36] R. Cowie, H. Gunes, G. McKeown, L. Vaclavu-Schneider, J. Armstrong, and E. Douglas-Cowie, "The Emotional and Communicative Significance of Head Nods and Shakes in a Naturalistic Database," *Proc. LREC Int'l Workshop Emotion,* pp. 42-46, 2010.

[37] P. Ekman and W.V. Friesen, *Facial Action Coding System.* Consulting Psychologists Press, 1978.

[38] R. Cowie and G. McKeown, "Statistical Analysis of Data from Initial Labelled Database and Recommendations for an Economical Coding Scheme," http://www.semaine-project.eu/, 2010.

[39] B. Jiang, M. Valstar, and M. Pantic, "Action Unit Detection Using Sparse Appearance Descriptors in Space-Time Video Volumes," *Proc. IEEE Int'l Conf. Face and Gesture Recognition,* 2011.

[40] H. Gunes and M. Pantic, "Dimensional Emotion Prediction from Spontaneous Head Gestures for Interaction with Sensitive Artificial Listeners," *Proc. Int'l Conf. Intelligent Virtual Agents,* pp. 371-377, 2010.

[41] H. Gunes and M. Pantic, "Automatic, Dimensional and Continuous Emotion Recognition," *Int'l J. Synthetic Emotions,* vol. 1, no. 1, pp. 68-99, 2010.

[42] H.G.M.A. Nicolaou and M. Pantic, "Automatic Segmentation of Spontaneous Data Using Dimensional Labels from Multiple Coders," *Proc. LREC Workshop Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality,* pp. 43-48, 2010.

[43] F. Eyben, M. Wollmer, M. Valstar, H. Gunes, B. Schuller, and M. Pantic, "String-Based Audiovisual Fusion of Behavioural Events for the Assessment of Dimensional Affect," *Proc. IEEE Int'l Conf. Face and Gesture Recognition,* 2011.

[44] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The Interspeech 2010 Paralinguistic Challenge," *Proc. 11th Ann. Conf. the Int'l Speech Comm. Assoc.,* pp. 2794-2797, 2010.

[45] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The First Facial Expression Recognition and Analysis Challenge," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* 2011.

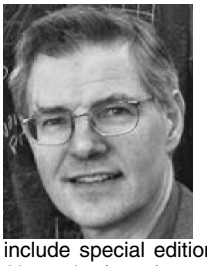
[46] K. Scherer and H. Ellgring, "Are Facial Expressions of Emotion Produced by Categorical Affect Programs or Dynamically Driven by Appraisal," *Emotion,* vol. 7, no. 1, pp. 113-130, 2007.

[47] K. Scherer and H. Ellgring, "Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns," *Emotion,* vol. 7, no. 1, pp. 158-171, 2007.

[48] R. Cowie, "Perceiving Emotion: Towards a Realistic Understanding of the Task," *Philosophical Trans. Royal Soc. London B,* vol. 364, no. 1535, pp. 3515-3525, 2009.

Gary McKeown received the PhD degree for research that explored mechanisms of implicit learning in the control of complex systems. He is a cognitive psychologist at the School of Psychology, Queen's University Belfast. His research focuses on communication, with interest in risk perception and decision making in environmental and health settings. This led to an interest in emotion and, in particular, the interrelationship of cognition and emotion. His recent research has focused on cross-cultural emotion perception and social signal processing.

Michel Valstar received the master's degree in electrical engineering from Delft University of Technology in 2005, and the PhD degree from Imperial College London in 2008. Both his master's and PhD theses were on the automatic recognition of facial expressions from face video. He is currently affiliated as a research associate with the intelligent Behaviour Understanding Group (iBUG) at Imperial College London. His research interests are in computer vision and pattern recognition, focusing on human sensing applications. He is a member of the IEEE.



Roddy Cowie is a professor of psychology at Queen's University, Belfast. He has used computational methods to study a range of complex perceptual phenomena—perceiving pictures, the experience of deafness, what speech conveys about the speaker, and, in a series of EC projects, the perception of emotion, where he has developed methods of measuring perceived emotion and inducing emotionally colored interactions. Key outputs include special editions of *Speech Communication* (2003) and *Neural Networks* (2005), and the *HUMAINE Handbook on Emotion-Oriented Systems* (2011). He is a member of the IEEE.

Maja Pantic received the PhD degree in computer science from Delft University of Technology in 2001. She is a professor of affective and behavioral computing at both the University of Twente and Imperial College London, where she heads the intelligent Behaviour Understanding Group (iBUG). She is the editor-in-chief of *Image and Vision Computing* and an associate editor for the *IEEE Transactions on Systems, Man, and Cybernetics Part B*, and has been a guest editor, organizer, and committee member for many major journals and conferences. Her research interests include computer vision and machine learning in face and body gesture recognition, multimodal human behavior analysis, and context-sensitive human-computer interaction. She is a fellow of the IEEE.

Marc Schröder is a senior researcher at DFKI, where he is a leader of the speech group and is responsible for building up technology and research in TTS. Within the FP6 NoE HUMAINE, he built up the portal http://emotion-research.net, which won the Grand Prize for the best IST project website 2006. He is an editor of the W3C Emotion Markup Language specification, project leader of the national-funded project PAVOQUE, and coordinated the FP7 STREP SEMAINE. He is an author of more than 65 scientific publications and has been a program committee member for many conferences and workshops.