The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent

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

Abstract—SEMAINE has created a large audiovisual database as part of an iterative approach to building agents that can engage a person in a sustained, emotionally coloured conversation, using the Sensitive Artificial Listener (SAL) paradigm. Data used to build the system came from interactions between users and an 'operator' simulating a SAL agent, in different configurations: Solid SAL (designed so that operators displayed appropriate non-verbal behaviour) and Semiautomatic SAL (designed so that users' experience approximated interacting with a machine). Having built the system, we recorded user interactions with the most communicatively competent version and baseline versions with reduced nonverbal skills. High quality recording is 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 labelled 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 to the scientific community through a web-accessible database.

Index Terms—Emotional Corpora, Affective Annotation, Affective Computing, Social Signal Processing

1 INTRODUCTION

IDEAS about the databases that emotion-oriented computing needs, have evolved with the discipline. Early research used archetypal expressions of discrete emotion categories. That was followed by an emphasis on naturalistic data [4] [**?**]. The balance has continued to shift, with a growing sense that data needs to be collected in a situation that is as close as possible to the one where the system will be used [**?**] [**?**]. The reason is that emotion is inherently interactive, and so the states that arise in a given situation, and the signs associated with them, are a function of the interactions that take place there [**?**].

This paper describes databases developed in a research effort which has followed that logic over a long period. The aim of the research has been to develop systems capable of holding a fluent, emotionally coloured conversation with a human being. Following the logic outlined above, that led to an iterative process where the desired system was simulated as well as current resources allowed; the results were used to simulate it better; and so on.

- *Gary McKeown and Roddy Cowie are with the School of Psychology, Queen's University Belfast, UK.*
- *Michel Valstar and Maja Pantic are with the Department of Computing, Imperial College London, UK. Maja Pantic is also with the University of Twente, The Netherlands.*
- *Marc Schröder is with DFKI GmbH, Saarbrücken, Germany.*

The data described here come from the SEMAINE project (Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression). It was directed towards a key intermediate stage: building a system that could engage a person in a conversation which was sustained, and emotionally coloured, but where there was no meaningful exchange of information. The system, called automatic SAL, now exists, and has been described elsewhere [8]. This paper describes the datasets that were generated in order to build the system, and have been generated through interaction with it.

The SEMAINE material represents a major shift from what has been available hitherto, mainly because it was developed specifically to address the task of achieving emotion-rich interaction with an automatic agent. Other resources have been recruited to that task, but they were rarely designed for it, and so they present various difficulties that the SEMAINE material avoids. Unlike the AMI meeting database, the SEMAINE material is rich in emotion [**?**]. In contrast to the various databases of acted material [**?**] [**?**] [**?**] [**?**] the emotion arises spontaneously from an activity. In contrast to sources that involve watching a film or undertaking a challenge [**?**] [**?**] the activity is conversation. In contrast to most databases derived from TV [**?**] [**?**] [7], there is information on both parties to the conversation. Rather than the short extracts which are commonly presented, the units are long enough to detect temporally extended patterns. Resources such as the Green Persuasive data [**?**]

or Canal9 [?] (which meet other criteria relatively well) show people interacting with a human, not a limited machine-like agent. An important consequence is that, unlike the SEMAINE material, they do not show how people behave when the agent mishandles the interaction which, for the foreseeable future, is likely to be an important issue.

In terms of technical specifications, the SEMAINE material is audio-visual, contrasting with the purely audio AIBO database [?] (which in many ways is SEMAINE's closest relative), and many others which are wholly or mainly visual [?] [?]. The recording quality in both modalities is high (unlike [?]), and the quantity is substantial (unlike [?]) the database currently contains over 45 hours of material. The annotation is very rich. It shows moment-by-moment variation in ratings (unlike [?] or [?]). The ratings cover emotion dimensions, emotion-related categories, and communicative categories making its cover even wider than the HUMAINE database [?]. Additional annotations pick out features that are of general interest – selected Action Units (AU), laughs, nods and shakes. Last but not least, it is available to the research community from the SEMAINE database website http://semaine-db.eu/.

The point of the comparisons with other resources is not to devalue them. It is to underscore the fact that if we want to achieve emotion-rich interaction with an automatic agent (which is presumably central to affective computing), then specific kinds of database are likely to be needed. The material described here represents the most extended effort so far to provide that kind of database.

2 THE SAL SCENARIO

SEMAINE is based on a scenario known as the 'Sensitive Artificial Listener', or SAL for short. It is described here briefly for completeness. More detail can be found in [2]. The scenario was designed to generate conversations where there is an abundance of the nonverbal features that sustain conversation and signal emotion, but where one party needs very little competence with spoken language. The point of defining that kind of role is that it is possible to build systems which adopt it, and therefore to record people interacting with them.

The interactions involve two parties, a 'user' (who is always human) and an 'operator' (either a machine or a person simulating a machine). What allows the operator to function with minimal spoken language competence is a 'script' composed of phrases that have two key qualities. One is low sensitivity to preceding verbal context: that is, the words that the user has just said do not dictate whether a given phrase can be used as the next 'move' in a conversation (though the way they were said may). The other is conduciveness: that is, the user is likely to respond to the phrase by continuing the conversation rather than closing it down. If an operator has a repertoire of phrases like that, he/she can conduct a conversation with quite minimal understanding of speech content. The idea was suggested by situations where humans seem to do something rather similar. For example, TV chat show hosts use stock phrases to draw out guests; and partygoers adopt a broadly similar strategy where the noise level makes it much easier to catch the nonverbal signals being given than the associated verbal content.

SEMAINE drew on scripts that earlier projects had refined iteratively. A key refinement was recognising that conversation tends to break down unless the operator appears to have a coherent personality and agenda. Given that the operator's communicative skills centre 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; Prudence is sensible, and 'tries' to make the user sensible; and Obadiah is gloomy, and 'tries' to make the user gloomy. That provides a simple way to create enough coherence and direction to keep users engaged, often quite emotionally and for quite long periods.

In effect, the SAL scripts provide a skeleton that can be fleshed out with non-verbal skills. The process of fleshing them out is iterative: data obtained from earlier versions of the scenario underpin development of the next version. The sequence of data collection is now described.

3 PHASES OF DEVELOPMENT

Iteration is fundamental to the style of data collection being presented here, and hence it is important to see how SEMAINE built on data collection in earlier projects.

The first systematic SAL recordings used 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. For instance, if the operator was simulating the Poppy character, and the user's mood was positive, the slide would show phrases that approved and encouraged happiness, accompanied by buttons which allowed the operator to change slides. For instance, if the user became angry, clicking a button would bring up a new slide, which displayed 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.

Two main sets of recordings were made with Powerpoint SAL, responding to requests for different kinds

TABLE 1
Powerpoint SAL recordings

Users	Sessions/User	Total time	Annotators
SAL 0	80/20	1:45	2
SAL 1	32/4	3:00	4
Hebrew SAL	20/5	2:30	—

of training data. One showed 20 users, each having a relatively brief interaction. The second showed four users, each having two, more sustained interactions. The first deliberately kept recording arrangements (audio and video) as unobtrusive as possible. The second used closer camera positions, brighter lighting, and head-mounted microphones. During both recording sets, both the operator and the users were in the same room.

Both bodies of data were annotated using the FEELtrace system [**?**], which allows raters to record their impressions of users' emotions in terms of the two most widely used 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).

Alongside these recordings, Powerpoint SAL was translated into Greek and Hebrew, and recordings were made with speakers of those languages. Table 1 summarises the material that is available from the Powerpoint SAL recordings.

Work with the Powerpoint system confirmed that users could have quite intense, sustained interactions with an operator whose conversation consisted entirely of phrases from a SAL-type script. It also indicated where the scripts needed to be revised, usually by adding responses that would have allowed a conversation to continue more naturally if they had been available. On that basis, the SEMAINE project was able to embark on a much more sophisticated program of data collection.

4 SAL SCENARIOS FOR SEMAINE RECORDINGS

SEMAINE recordings contrast with earlier SAL material at several levels. Recording quality was much higher (see section 5.2). It was much easier for the user to regard the operator 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 pre-defined list but (unlike Powerpoint SAL) the system speaks them; and Automatic SAL, where an automated system chooses sentences and non-verbal signals. These were chosen to 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 allow different behaviours to be observed.

Generally speaking, the three scenarios were recorded using disjoint sets of participants. Only a few participants participated in more than one scenario, and none participated in all three.

4.1 Solid SAL

The Solid SAL scenario was designed to record behaviours (mainly non-verbal) associated with relatively strong engagement between user and operator. In particular, it was designed to capture a range of nonverbal behaviours that are part of a normal conversation – backchannelling, eye contact, various synchronies, and so on. That kind of engagement is difficult to achieve 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 use the exact phrases in a SAL script.

Users were encouraged to interact with the characters as naturally 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 less like machine human interaction than the other scenarios, 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.

In total 24 recording sessions used the Solid SAL scenario, recordings were made of both the user and the operator and there were approximately 4 character interactions in each recording sessions, providing a total of 95 character interactions and 190 video clips to the database.

4.2 Semi-Automatic SAL

Semi-automatic SAL was similar to powerpoint SAL in that a human operator chose between a number of stock phrases. These were made available to her/him through a Graphical User Interface which is illustrated in Fig. 1. The selected phrase was then played using a pre-recorded audio file spoken by an actor whose voice had been judged appropriate for the character. As well as hearing the voice, the user saw a simplified screen designed to keep attention focussed in the general direction of the camera. It is illustrated in Fig. 1. The

feature designed to hold attention was the 'mouth' of an abstract face, which was formed by the spectrum of the speech. 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 audio either switched off (Experiment 2); or with audio filtered to remove verbal information (Experiment 3). The degradation made it harder for the operator to avoid inappropriate choices of the kind that the automatic system would necessarily make (because it does not use linguistic information), and resulted in recordings where users showed various signs of communication breakdown.

In the degraded versions of Semi-automatic SAL, one of the four character interactions was with the full Semi-automatic SAL system while the other three were degraded. 11 Semi-automatic SAL recording sessions took place in a manner directly comparable to Solid SAL. A further 25 sessions took place with differing degrees of degradation of information to the operator. Only the user video is made available in the database. The 4 character interactions for each recording session in the Semi-Automatic SAL experiments add a further 144 videos to the database. See Table 3 for an overview.

4.3 Automatic SAL

In the fully automatic SAL recordings, the utterances and non-verbal actions executed by the SAL Character were decided entirely automatically by the current version of the SEMAINE project system [8]. The SAL characters are represented using life-like avatars whose appearance is stereotypical for their character (see Fig. 2), and they speak with a synthetic unit selection voice, which again is stereotypical. Using the greyscale camera aimed at the user, the SEMAINE system detected when a person's face was present, whether the person nodded or shaked their head, and whether the person smiled, raised their eyebrows, lowered their eyebrows, or opened their mouth. The detected head nods and shakes were used to predict the emotional state of the user in terms of a 5 dimensional descriptor (the fully rated dimensions described in section 6.1.1). Using the head-mounted microphone, the system identified whether the user was speaking or not, gave an indication whether the participant's speech turn had finished, detected a number of key words (including the characters' names), and predicted the emotional state of the user. A dialogue manager keeps track of the flow of the conversation, as well as the nonverbal communicative acts of the user and her/his emotion, to decide what to say next. It also decides whether any non-verbal communicative acts should be performed by the avatar.

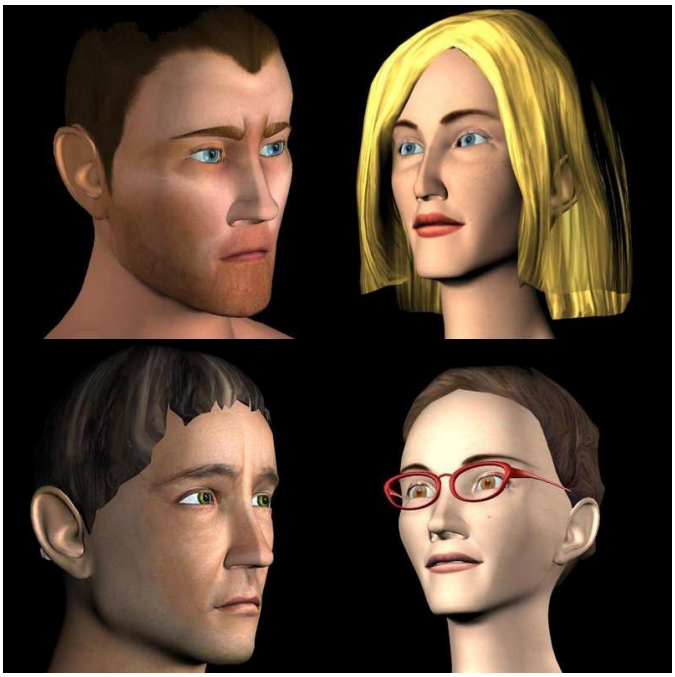

Fig. 2. The four SAL character avatars. Clockwise from top-left: Spike, Poppy, Prudence and Obadiah.

Participants interact with two versions of the system meaning they interact with each of the four characters twice. Sessions are limited to approximately 3 minutes or if the participant does not engage with the system they are cut short after a minimum of 1.5 minutes. At time of print there have been 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. An initial experiment compared a version of the system based on SEMAINE system 3.0.1 (revision 734) with a degraded System in which visual feedback was turned off and user emotional state was randomly chosen, 15 participants were tested with this configuration adding 30 recording sessions to the database. A second experiment used two different system versions; the full version was based on SEMAINE system 3.0.1 (revision 753) while the degraded system removed most of the affective cues from the system leaving only a stark basic SAL scenario with no backchanneling, emotional information and random utterance selection and flat affect in the agent voices. 20 participants were tested with this configuration adding a further 40 recording sessions to the database. A third experiment used a different full version of the system based on SEMAINE system 3.0.1 (revision 782) which featured improved dialogue management and was compared with the same degraded system used in the second experiment. At time of print we have 13 participants tested with this configuration adding a

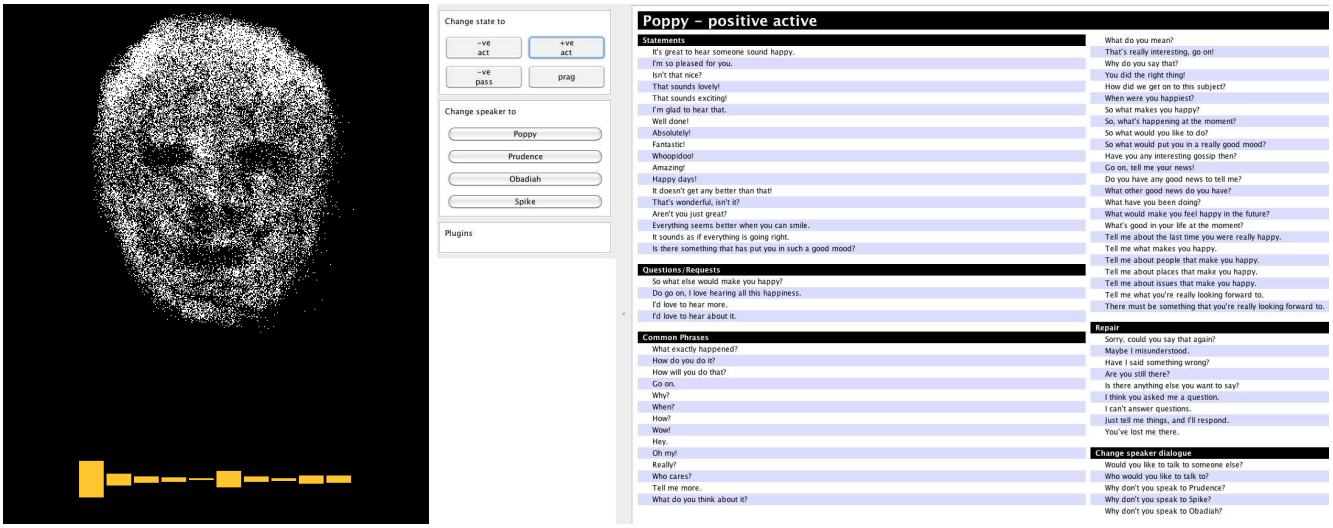

Fig. 1. Semi-Automatic SAL screens for the user and the operator

further 36 recording sessions. We project that the number of recordings in Automatic SAL experiment will be in the region of 90 participants making a total of 180 recording sessions with various versions of the system; these will be divided into character interactions providing another 720 videos to the database.

5 Procedural Specifications

5.1 Participants and procedure

Participants were undergraduate and postgraduate students. 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 thirty minutes with an approximate interaction time of five minutes per character, though there were considerable individual variations. Participants were told to change 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 one hour with eight character interactions of approximately three 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 room and recording starts. The operator/agent recited a brief introduction script and the interaction began.

Once the interactions were complete a debriefing session took place, allowing the user to ask more about the system.

5.2 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. These considerations guided the decisions on the choice of sensors, and how the sensors are placed.

Sensors. Video is recorded at 49.979 frames per second and at a spatial resolution of 780 x 580 pixels using AVT Stingray cameras. 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. 4.

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 [9].

To record what the User and the Operator are saying, we use two microphones per person recorded: 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. The wearable microphones are AKG HC-577-L condenser microphones, while the room microphones are AKG C1000-S microphones. This results in a total of four micro-

Fig. 3. Images of the recording setup for both the User (left) and Operator (right) Rooms.

Fig. 4. 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 3214th frame of the 19th recording.

phones 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. Audio is recorded at 48 kHz and 24 bits per sample.

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 through a set of teleprompters. Within each teleprompter, the cameras recording a person's frontal view 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. Because the target application of the SEMAINE system would be interaction with a conversational agent in controlled conditions, we used professional lighting to ensure an even illumination of the faces. Images of the two rooms can be seen in Fig. 3.

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. [10]. This system 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 us to synchronise the audio and the video sensor data with a maximum time difference between data samples of 25 microseconds.

TABLE 2
Solid SAL recordings

Users	Sessions/User	Total Time	Annotators
Solid SAL User	95/24	480	6+
Solid SAL Operator	95/4	480	1

Data compression The amount of raw data generated by the visual sensors is very high: 959 character interactions, lasting on average 5 minutes, recorded at 49.979 frames/second at a temporal resolution of 780*580 pixels with 8 bits per pixel for 5 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 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.

5.3 Summary of the SEMAINE Recordings

Tables 2, 3 and 4 summarise the recordings that make up the database.

6 Annotation & Associated Information

6.1 FEELTrace annotation

In Solid SAL and Semi-automatic SAL trace-style continuous ratings were used to record how raters perceived

TABLE 3
Semiautomatic SAL Recordings. Time is measured in minutes.

Experiment & System	Sessions/User	Approximate Total Time	Annotators
Experiment 1			
Full audio	44/11	345	1
Experiment 2			
Full audio	22/11	345	1
No Audio	22/11	125	1
Experiment 3			
Full audio	24/12	345	1
Degraded Audio	12/12	125	1
Degraded Audio & No Vision	12/12	125	1

TABLE 4
Automatic SAL recordings. Time is measured in minutes.

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	48/12	144	1
Degraded 2	48/12	144	1
Pilots	20/5	60	1

users' emotions and other related states. Note that perceived emotion is what the system needs to know about: it should respond as a person would, even if the person would be wrong. [4]

The labels were chosen from a 'wishlist' of annotations produced by consultation amongst SEMAINE members. These included classic dimensions of emotion and a body of categorical labels that may be present in only some of the clips. The details of the items chosen for annotation follow.

6.1.1 Dimensions

The rating procedure involved full rating for five dimensions and then optional rating for instances of another 27 dimensions. The five fully rated dimensions are:

- Valence
- Activation
- Power
- Anticipation/Expectation
- Intensity

These are all well established in the psychological literature. An influential recent study [11] argues that the first four Valence, Activation, Power and Expectation account for most of the distinctions between everyday emotion categories. Valence is an individual's overall sense of 'weal or woe': does it appear that on balance, the person rated feels positive or negative about the things, people, or situations at the focus of his/her emotional state? Activation is the individual's global feeling of dynamism or lethargy. It subsumes mental activity as well as physical, preparedness to act as well as overt activity. The power dimension subsumes two related concepts, power and control. However, people's sense of their own power is the central issue that emotion is about, and that is relative to what they are facing. Anticipation/Expectation also subsumes various concepts that can be separated – expecting, anticipating, being taken unawares. Again, they point to a dimension that people find intuitively meaningful, 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 not something that should be assumed.

The other traces dealt with more or less categorical descriptions, and were made after the five core dimensions have been annotated. After rating the clip on five dimensions, the rater was thoroughly familiar with its contents. He/she was then presented with a list of emotion- and communication-related categories, and 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 option was very rarely used. The items fell into four broad groups.

Basic Emotions: 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 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. [**?**]). Hence the labels in this group are:

- Fear
- Anger
- Happiness
- Sadness
- Disgust
- Contempt
- Amusement

Epistemic states: These states were highlighted by Baron-Cohen et al [**?**], and have a roused a lot of interest in the machine perception community. They are relatively self-explanatory. As before, they are labelled where the clip contains a relatively clear-cut example of the state in question. For example, the guidelines suggest using the category 'certain/not certain' if there is an episode where the fact that someone is certain about something stands out; but it should not be selected simply because the person seems to accept something unquestioningly (e.g. that the sun will rise tomorrow).

The labels included in this section are:

- Certain / not certain
- Agreeing / not agreeing
- Interested / not interested
- At ease / not at ease
- Thoughtful / not thoughtful
- Concentrating / not concentrating

Interaction Process Analysis: These labels are of particular use in dialogue management. They are a subset of the system of categories used in Interaction Process Analysis [**?**]. The intention is not to provide a full interaction process analysis, but to indicate when the issues that it highlights become salient. The labels included in this section are:

- 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 aims to highlight cases where there user is not communicating his or her feelings in a straightforward way. Among other things, that means that the material should be used carefully or not at all in a training context. The included labels are:

- 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. [**?**] 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.

TABLE 5
Amount of Labelling

Users	Type of Labelling	Labellers
Solid SAL	Full	6+
Semi-automatic SAL	Full	1+
Automatic SAL	Engagement	1

6.1.2 Amount of Annotation

The amount of annotation differs between the experiments. Solid SAL was completed first and has the largest body of annotation. Semi-Automatic SAL has a smaller subset and Automatic SAL has the least annotation.

Solid SAL: The user video clips have the most annotation. There are trace ratings by at least one annotator for all the Solid SAL user clips. 17 have been annotated by at least 6 raters and 4 clips have been annotated by 8 raters. These annotations include the five core dimensions and four optional categories. A substantial proportion of the operator clips have been annotated by at least one rater and three clips have been annotated by three raters.

Semi-Automatic SAL: A smaller amount of annotation is included in the database for the Semi-Automatic SAL clips. Annotations by two raters have been completed for some of the Semi-Automatic SAL clips and a substantial proportion of the operator clips have been annotated by one rater. These are again the five core dimensions and four optional categories annotations.

Automatic SAL: The least amount of annotation is provided for the Automatic SAL clips. Due to the SEMAINE project time constraints the only trace style annotations possible are continuous traces that are recorded live as the interaction is recorded, these are traced by one rater watching a live video feed of the interaction. The dimension chosen for these ratings was engagement in the conversation as this was the most suitable for the evaluation needs of the project.

Table 5 summarises the amount of labelling that each type of recording has received. It is expected that annotation will be extended gradually.

6.2 Transcripts

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

6.3 Interaction evaluations

It is a feature of Semiautomatic and Automatic SAL sessions that the interaction sometimes breaks down.

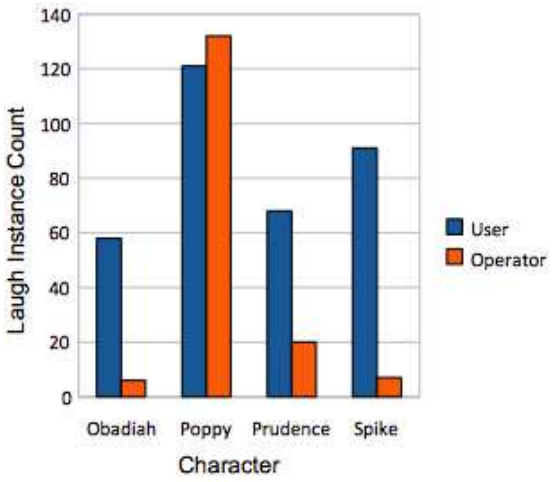

Fig. 5. Instances of user and operator laughter for each character in Solid SAL recordings 1-21.

Hence they provide an opportunity to study the signals of breakdown. Several kinds of data are available to identify sessions where problems arose. The experimental procedure in Semiautomatic and Automatic SAL included three questions 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?"; "How much did you feel you were involved in the conversation?". The sessions also included a 'Yuck button', which users were asked to press when the interaction felt unnatural or awkward. In both Semiautomatic 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.

6.4 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 individually identifiable instances of laughter for each character for both user and operator can be seen in Figure 5.

6.5 Nods and Shakes

Work with the teams at Imperial College London and Queen's University Belfast sought to refine knowledge concerning Nods and Shakes from a subset of nods and shakes drawn from the SEMAINE database. 154 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 a set of annotations derived from McClave [**?**] 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 these annotations were subjected to a cluster analysis and the results and greater detail regarding the annotations can be found in [**?**].

6.6 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 [**?**]. Instances were labelled 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. These Action Unit annotations will be made available with the SEMAINE database.

7 CHARACTERISTICS OF SOLID SAL ANNOTATIONS

7.1 Reliability of main traces

Reliability was 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 alpha measure of agreement can be calculated. Table 6 summarises the results. Overall, the findings confirm that most of the ratings are reliable in some broad 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 that raters agree. For valence, it is their frequency.

It is more difficult to measure intra-clip agreement (that is, agreement between raters on the way a single measure, say valence, rises and falls in the course of a single clip). It is possible to calculate a correlation between the list of values that defines one rater's trace of valence for a target clip and the list that defines another

TABLE 6
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 non-acceptable values)

	Intensity	Valence	Activation	Power	Expectation
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 7
Reliability Analysis

	QA analysis	Correlational (α) analysis
Total no. of datasets (i.e. sets of 6 or 8 traces of a particular clip on a particular dimension)	305	303
Fail stringent test (alpha>0.85, p(QAg)<0.01)	90	104
Fail moderate test (alpha>0.75, p(QAg)<0.05)	43	41
Fail minimal test (alpha>0.7)	n/a	28

rater's; and alpha coefficients can be derived from those correlations. However, there are reasons to be wary of correlation as a measure: successive points in a trace are not independent, and it may be more appropriate to consider consider measurement as ordinal rather than interval. A new method of calculating agreement which avoids these problems has been developed. It is called QA, for Quantitative Agreement, and it is described in [**?**] Table 7 summarises the results. It can be seen that 2/3 of the traces pass a stringent test in terms of either criterion, and about 80% reach a level that would normally be regarded as acceptable. Overall, the QA measure is slightly more stringent, so that the numbers reaching the standard criterion on it ($p<0.05$) are comparable to those reaching an alpha of 0.75. Less than 10% fail to reach the standard criterion of alpha=0.7.

The overall picture masks some differences between the different types of trace. The long established affective dimensions, intensity, valence, and activation are substantially more likely to pass the stringent QA test than the stringent alpha test. The order among intensity, valence, and activation traces is as would be expected from history (and the subjective ease of the rating). Expectation behaves differently: only half of the expectation traces pass the stringent QA test, whereas 6/7 pass the stringent alpha test. These differences presumably reflect differences in the psychological status of the different scales, and they deserve to be followed up. Nevertheless, the overall picture is very positive. Whichever measure is used, the great majority of the datasets for all trace types reach what would normally be regarded as an acceptable level.

7.2 Distribution of optional traces

The 'optional' trace categories identify regions where raters felt that particular qualitative descriptors applied, and show how the chosen states appeared to change over time. Table 8 provides an overview of the choices, showing how often each option was used. Table 9 shows the distribution of the most frequently used optional traces for each of the characters (for the sake of balance, only data from the six raters who traced all the clips is 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.

Table 9 shows 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 fact that so many of the options are used so little in itself indicates a considerable measure of agreement among raters.

8 Automatic Analysis of the Database

On one hand the SEMAINE corpus is a valuable repository to study the SAL paradigm, and in more general terms the way a human might interact with artificially intelligent agents. On the other hand, it provides an excellent opportunity to develop new ways of automatically analysing human behaviour 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. Besides acting as the main resource to develop the SEMAINE system [8], the SEMAINE database has

TABLE 8
Solid SAL Additional Category Annotations (numbers represent the raw number of annotations for each dimension at time of print).

Basic Emotions		Epistemic States		Interaction Process Analysis		Validity	
41	Anger	78	(not) certain	11	Shows solidarity	27	Breakdown of engagement
7	Disgust	213	(dis) agreement	29	Shows Antagonism	5	Anomalous Simulation
172	Amusement	39	(un) interested	25	Shows tension	21	Marked sociable Concealment
93	Happiness	101	(not) at ease	18	Releases tension	10	Marked sociable simulation
58	Sadness	109	(not) thoughtful	23	Makes suggestion		
25	Contempt	27	(not) concentrating	5	Asks for suggestion		
3	Fear			147	Gives Opinion		
				10	Asks for opinion		
				220	Gives information		
				12	Asks for information		

TABLE 9
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

already been used successfully for a number of other related projects.

Jiang et al. [**?**] 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 Quantisation), and found that between the two Local Phase Quantisation performed best. They were able to detect 7 AUs with an average F1-measure of 76.5%. However, this was tested on only 8 sessions of only two subjects. The authors found that there was a big difference in performance between the two subjects, making it hard to assess how well their system performs on this type of data. They also reported on the detection of AUs on posed facial expressions, where they were able to leverage the temporal information contained in the continuous AU coding of that data. They reported that the temporal extension of LPQ, called LPQ-TOP, attained the highest performance.

Gunes and Pantic [**?**] 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 5 dimensions labelled in all recordings (arousal, expectation, intensity, power, and valence). Using ratings of multiple observers annotating a subjective phenomena is notoriously difficult [**?**]. To deal with the problem of differences in interpretation by different observers, they chose to model each annotator directly, independent of the others. Thus, a separate Support Vector Regression model was trained for each observer, and the predictions of those models were compared with the annotations provided by the same observer.

Nicolaou et al. [**?**] 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 maximising inter-coder agreement, identify transitions between emotional states, and that automatically segment audio-visual data so it can be used by machine learning techniques that require pre-segmented sequences. They tested their approach on the SEMAINE corpus and reported that the segmentation process appeared to work as desired, with the segments identified by their algorithm capturing the targeted emotional transitions well.

Eyben et al. [**?**] used the SEMAINE corpus to first detect a range of non-verbal 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, as they occurred very frequently in the SEMAINE data. The detected events were detected on the basis of a short

temporal window (approximately half a second), and combined into a single bag-of-words feature vector. A comparison was made with a signal-based approach, where the audio and video features originally extracted to detect the events were instead used directly to detect the dimensional labels. 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.

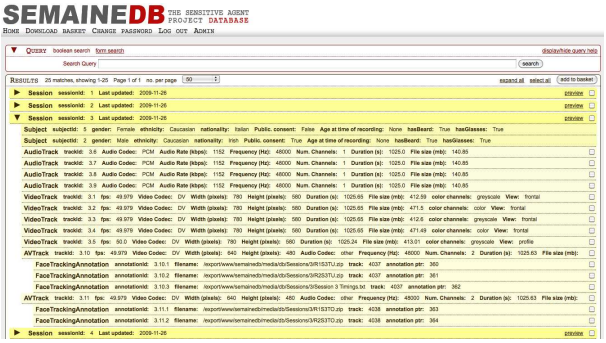

Fig. 6. Data organisation of the database.

9 AVAILABILITY

The SEMAINE Solid-SAL dataset is made freely available to the research community. It is available through a web-accessible interface with url http://semaine-db.eu/. The available dataset consists of 150 recordings, featuring as many participants. Some of the participants play the role of the Operator in a session, but they also appear in the User role in some of the other interactions. The youngest participant was 22, the oldest 60, and the average age is 32.8 years old (std. 11.9), 38% are male. Although the participants come from 8 different countries, almost all of the participants come from a Caucasian background.

9.1 Organisation

Within the database, the data is organised in units that we call a *Session*. In a Session, which is part of a recording, 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 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 consist of 9 sensors, while all other scenarios have 7 sensors associated with them. 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, 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.

Approximately one-third of the recorded data is being withheld from public access to allow for benchmarks procedures to be set up and for the organisation of challenges similar to the Interspeech audio-analysis series (e.g. [**?**]) and the FERA facial expression recognition challenge [**?**]. The database also defines a partitioning of the publicly available data into a training and a development set. The former would be used by researchers to train their systems with all relevant parameters set to a specific value, while the latter 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 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

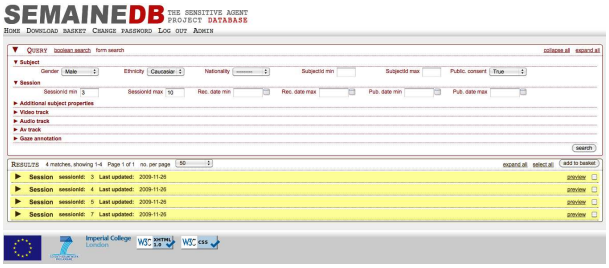

Fig. 7. Form search: some options and the search results.

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.

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. For instance, the Solid SAL data is labelled with five affective dimensions, as opposed to the more usual two. Recognising the less common dimensions on the basis of face, voice and text is a natural challenge. Similarly, there is information about the user's level of engagement in both Automatic and Semi-automatic SAL. Recognising level of engagement is a natural challenge, and probably not too intractable.

Beyond that, the quality of the data that is available makes it natural to add new types of information. That has various levels. It would make sense to extend the kind of tracing that has been applied to Solid SAL 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. The most obvious types of gesture have already been identified – 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 conversational elements, and their relationship to the parties' emotions and engagement.

These developments have clear applications in computing, but they would also support advances in the human sciences. For example, substantial theoretical issues hinge on the way facial gestures appear in spontaneous emotional expression; but the scarcity of naturalistic material, and the labour of identifying facial actions, has made it difficult to draw strong conclusions [?] [?]. The issue affects not only the generation of emotion-related signals, but also the mechanisms needed to recover information from such signal configurations [?]. SAL data offers one of the most promising ways to address these questions.

A deeper theoretical question hinges on the points which has been emphasised throughout, which is that interacting with an artificial agent is not the same as interacting with a human. The superficial response is to treat them as separate problems. The deeper response is to use the contrast as a way to expose the multitude of factors that make human-human interaction what it is, but whose effect is usually so automatic that we do not realise they are there.

Last but not least, the SEMAINE approach to data collection provides a model that it makes sense to generalise. 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 semi-automatic simulations of the SAL dialogue system through 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 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. D.-C. 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] ——, "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] M. Schröder and et al., "Building autonomous sensitive artificial listeners," *Transactions on Affective Computing*, 2010, under revision.

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

[10] J. Lichtenauer, J. Shen, M. 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.

[11] 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, Feb 2007.

Gary McKeown is a cognitive psychologist at the School of Psychology, Queen's University Belfast. His PhD explored mechanisms of implicit learning in the control of complex systems. 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 inter-relationship of cognition and emotion. Recent research has focused on emotion, social signal processing and the cross-cultural emotion perception.

Michel Valstar received his masters in Electrical Engineering at Delft University of Technology in 2005, and his Ph.D. from Imperial College London in 2008. Both his masters 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 techniques, focusing on human sensing applications.

Roddy Cowie studied Philosophy and Psychology as an undergraduate, and received his PhD from Sussex on relationships between human and machine vision. He joined the psychology department at Queen's, Belfast in 1975 and became professor in 2003. He has applied computational methods to the study of complex perceptual phenomena in a range of areas – perceiving pictures, the subjective experience of deafness, the information that speech conveys about the speaker, and most recently the perception of emotion through a series of EC projects. He developed influential methods of measuring perceived emotional colouring and inducing emotionally coloured interactions. He has authored or edited several landmark publications in the area, including special editions of Speech Communication (2003) and Neural Networks (2005), and the HUMAINE handbook on emotion-oriented computing (2010).

Maja Pantic received the MSc and PhD degrees in computer science from Delft University of Technology, The Netherlands, in 1997 and 2001, respectively. She is a professor of affective and behavioural computing at both the Computer Science Department of the University of Twente and the Computing Department of Imperial College London, where she heads the intelligent Behaviour Understanding Group (iBUG). She is the editor-in-chief of the Image and Vision Computing Journal and an associate editor for the IEEE Transactions on Systems, Man, and Cybernetics Part B. She is a guest editor, organizer, and committee member for more than 10 major journals and conferences. Her research interests include computer vision and machine learning applied to face and body gesture recognition, multimodal human behavior analysis, and context-sensitive humancomputer interaction (HCI). She is a senior member of the IEEE.

Marc Schröder is a Senior Researcher at DFKI and the leader of the DFKI speech group. Since 1998, he is responsible at DFKI for building up technology and research in TTS. Within the FP6 NoE HUMAINE, Schrder has built up the scientific portal http://emotion-research.net which won the Grand Prize for the best IST project website 2006. He is Editor of the W3C Emotion Markup Language specification, Coordinator of the FP7 STREP SEMAINE, and project leader of the national-funded basic research project PAVOQUE. Schrder is an author of more than 65 scientific publications and PC member in many conferences and workshops.