Building Autonomous Sensitive Artificial Listeners

Marc Schröder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Björn Schuller, Etienne de Sevin, Michel Valstar, and Martin Wöllmer

Abstract—This paper describes a substantial effort to build a real-time interactive multimodal dialogue system with a focus on emotional and non-verbal interaction capabilities. The work is motivated by the aim to provide technology with competences in perceiving and producing the emotional and non-verbal behaviours required to sustain a conversational dialogue. We present the Sensitive Artificial Listener (SAL) scenario as a setting which seems particularly suited for the study of emotional and non-verbal behaviour, since it requires only very limited verbal understanding on the part of the machine. This scenario allows us to concentrate on non-verbal capabilities without having to address at the same time the challenges of spoken language understanding, task modeling etc. We first report on three prototype versions of the SAL scenario, in which the behaviour of the Sensitive Artificial Listener characters was determined by a human operator. These prototypes served the purpose of verifying the effectiveness of the SAL scenario and allowed us to collect data required for building system components for analysing and synthesising the respective behaviours. We then describe the fully autonomous integrated real-time system we created, which combines incremental analysis of user behaviour, dialogue management, and synthesis of speaker and listener behaviour of a SAL character displayed as a virtual agent. We discuss principles that should underlie the evaluation of SAL-type systems. Since the system is designed for modularity and reuse, and since it is publicly available, the SAL system has potential as a joint research tool in the affective computing research community.

Index Terms—Embodied Conversational Agents, Rapport Agents, Emotion recognition, Emotion synthesis, Real-time dialogue, Listener behaviour, Turn-taking

1 Introduction

MAKING the interaction with computers more natural for humans requires computers to acquire multiple capabilities. Alongside many others, these capabilities include aspects of communication that are emotion-related and non-verbal. This paper describes a sustained effort to bring together, in one consistent framework, many of the concepts and technologies required to endow a computer with such capabilities, and describes how these concepts and technologies are put to use to implement a specific type of dialogue system: a fully autonomous implementation of 'Sensitive Artificial Listeners' (SAL).

We consider a one-to-one dialogue situation where, at any given time, one user is having a conversation with one virtual agent character. The interaction is multimodal, involving speech, head movements, and facial expressions. One of the key features of human interactions that we expect to reproduce is that the dialogue will involve some emotional colouring – not in terms of episodes of intense "basic" emotions, but in the sense of emotion "pervading everyday life" [1].

The goal of having a natural conversation in this setting sets technology a number of substantial challenges. The computer must be able to perceive the user's verbal and non-verbal behaviour, i.e. have some awareness of the words spoken, the prosody with which they are spoken, and the user's head movements and facial expressions. While the user is talking, the computer must exhibit suitable *listener behaviour* – notably multimodal *backchannels* [2] which signal that the listener is still present and following what is being said, and which may at the same time provide *feedback* to the speaker about the listener's reaction [3]. Examples of such listener behaviour are head nods, smiles, or short vocalisations such as "uh-huh" or "wow", which may be produced individually or in combination. The computer must determine when is a good moment to *take the turn* [4] and become the speaker itself; it must then produce a verbal utterance which must fit the dialogue context, and which has to be spoken with a suitable voice and synchronous facial animation, including lip movements, facial expressions and head movements. If any of these processes is not performed, or does not match the user's sense of natural behaviour, the quality of the interaction is degraded.

- *M. Schröder and S. Pammi are with German Research Centre for Artificial Intelligence (DFKI), Germany.*
- *R. Cowie and G. McKeown are with Queen's University, Belfast, UK.*
- *M. Pantic, and M. Valstar are with Imperial College of Science, Technology and Medicine, London, UK.*
- *H. Gunes is with EECS, Queen Mary University of London, UK.*
- *D. Heylen and M. ter Maat are with Universiteit Twente, Netherlands.*
- *B. Schuller, F. Eyben and M. Wöllmer are with Technische Universität, München, Germany.*
- *C. Pelachaud, E. Bevacqua and E. de Sevin are with CNRS–LTCI, Telecom ParisTech, Paris, France.*

The present paper reports on an effort to integrate these capabilities into one software architecture in order to provide the building blocks needed for studying natural one-to-one conversations between a virtual character and a human user.

It is important to notice that the requirements formulated above are quite distinct from the various requirements arising from the task-oriented interactions that are often studied. For example, an accurate interpretation of the user's words depends on high-quality Automatic Speech Recognition (ASR) [5]; efficiently achieving a dialogue goal depends on suitable dialogue structures and on modelling task domains [6]; and having a common experiential basis for an interaction requires grounding of a machine's knowledge about the world [7]. We recognise the importance of these goals, but they are not the goals that we address in this study. They represent one of the streams that need to converge to produce a competent artificial interactant; our work represent another, which has received much less attention in the computational community, but which psychological research suggests is at least equally important [8], [9].

The scenario that we call "Sensitive Artificial Listeners" was developed specifically to let us concentrate on the emotional and non-verbal competences involved in conversation. Its raison d'être is to avoid the difficulties of task-oriented dialogues and instead address directly the emotion-related and non-verbal capabilities that a system needs in order to have a naturally flowing conversation. A good deal of experience indicates that the two aspects can operate rather independently (for example, when a party is too noisy for people to hear most of each other's words, and yet interaction flourishes). If so, it makes sense to expect that capabilities developed in the SAL scenario can later be integrated into task-oriented interactions.

The purpose of the present article is to describe the 'big picture' of SAL development – the reasoning and development of the SAL scenario as such; its test-runs in various human-operated versions; and the key aspects of its implementation as an Autonomous SAL. Rather than describing the technical details of all components involved, which would not be possible in a single paper, we attempt to convey the essential functionality provided by the various components and a sense of how they co-operate to provide the intended functionality. In evaluating the system, we start from principles before describing the evaluation and results. Substantial background information is available for most aspects of the work described here; the reader is referred to the SEMAINE project website http://www.semaine-project.eu, and notably to the following publications: [10], [11], [12], [13], [14], [15], [16], [17].

The paper is structured as follows. We start by giving an account of the current state of the art in emotion-oriented and non-verbally sensitive dialogues between humans and computers (Section 2). We then motivate why we consider Sensitive Artificial Listeners to be a promising scenario (Section 3), and describe several human-driven versions of SAL which served to understand the relevant variables and to collect data for use in quantitative and qualitative analyses (Section 4). In the main part of the paper (Section 5), we provide an overview of the Autonomous SAL system, describing its design principles, architecture, and the capabilities and limitations of the individual system components. We discuss principles for evaluating systems such as ours before presenting an evaluation of the Autonomous SAL system (Section 6). The paper concludes with a discussion of the specific contribution made by the present work to the progression of the Affective Computing research area. An example in the Appendix shows the type of interaction that can occur with the Autonomous SAL system.

2 Related work

In the mid 1990's, spoken dialogue systems were investigated under the perspective of human language interfaces to information [18], [19]. Major concerns at the time were speech recognition and language understanding accuracy, and the challenge of extending the speech recognition vocabulary for open domains [18]. Interaction was modelled as a ping-pong scenario where user and system would speak in turn.

Conversational dialogue systems [20] improved on the dialogue capabilities by adding the notion of system goals and dialogue obligations, leading to richer mixed-initiative dialogues. The notion of incremental analysis of user behaviour is introduced and identified as a precondition for giving listener feedback in order to ground the degree of mutual understanding in the dialogue. Despite progress in this area, for example using machine learning methods to improve turn-taking behaviour [21], [22], much remains to be done. For example, it is only partially understood when to generate backchannel vocalisations [23] and how their meaning is interpreted [24].

Despite intense efforts, the single most important source of problems for spoken dialogue systems is still considered to be the limited accuracy of speech recognition technology [25]. Multimodal dialogue systems attempt to complement the verbal communication channel with other relevant modalities, including visual displays and the analysis of the user's non-verbal behaviour. A major example is the SmartKom system [26], which emphasised symmetric multimodality. SmartKom is a multimodal dialogue system in the information kiosk scenario, where the system's task is to provide information and functionality to the user. The system has sophisticated support for multiple input and output modalities. The user interacts with a small i-shaped information agent named "Smartakus". Non-verbal behaviour by the user, such as

deictic gestures, is interpreted in conjunction with speech input in order to resolve references. A typical scenario involves a user pointing to an area on a display and saying "I want to reserve this seat". In addition, the system has initial support for emotion recognition from speech and from facial expressions, and shows some limited expressivity to inform the user about the system's internal state while a system response is being generated.

Embodied Conversational Agent (ECA) interfaces [27] change the metaphor from interacting with a "system" to interacting with an "agent", which often takes the appearance of a human-like face or body displayed on a computer screen. ECA interfaces make it natural to reason in terms of human-human multimodal interaction capabilities.

On the one hand, the ECA can exhibit expressive facial and bodily behaviour [28] and speak with an expressive synthetic voice [29]. For example, the NECA project [30] generated scripted dialogues between ECAs in a social web community and a product showroom environment. The dialogue scripts contained annotations on how to emotionally influence the facial expression and tone of voice. The VirtualHuman project [31] supported dialogues involving multiple humans and multiple ECAs. Emotions to be expressed by the ECAs were computed by rules operating on the domain knowledge, and were realised through facial expression, skin texture, tears, and breathing patterns. In these scenarios, the ECAs' expressive behaviour is determined by the game or application logic. The user's non-verbal behaviour is not taken into account.

The expressive behaviour that is generated can have a positive effect on the dialogue. For example, Pardo et al. [25] found that the presence of an ECA showing contextually appropriate gestural behaviour had a positive effect on objective measures of dialogue success related to turn management and error recovery. At the same time, the perception of expressive behaviour appears to depend on user characteristics. For example, Krämer et al. [32] found systematic effects of the users' gender, age, and computer literacy on their assessment of an ECA's non-verbal behaviour. For example, female users gave better ratings than men when the agent showed more smiles and self-touching gestures; men preferred less non-verbal behaviour.

On the other hand, dialogue systems can also focus on the analysis of the user's non-verbal expressive behaviour without showing expressivity themselves. To do that, they rely on the analysis of the user's behaviour from face [33] and voice [34]. An example use case for this capability is an emotion-aware voice portal [35], which detects a customer's anger in time to redirect the customer to a human agent.

In ECA systems, the question of "social presence" becomes relevant: does the user perceive the ECA as a social entity, and interact according to social conventions? Nass and Moon [36] found people to show a tendency to treat computers as social entities, especially in states of "mindlessness", even though they would not ascribe any anthropomorphism to them. Achieving such a perception of 'social presence' depends, according to Biocca et al. [37], on co-presence, psychological involvement, and behavioural engagement. These criteria can be further characterised in terms of factors such as mutual awareness, mutual attention, and behavioural interaction [38]. In other words, for an ECA to be perceived as a social entity, it is important that it shows responsiveness – it must show some awareness of the user's non-verbal behaviour, and react to it. In doing so, it needs to take into account non-verbal user behaviour in dialogue planning [39], and generate non-verbal behaviour while speaking [40], [41] and while listening [42]. In order to be perceived as natural, the ECA's behaviour needs to follow the principles underlying the use of non-verbal behaviour in conversation [43].

Works focusing on responsiveness in human-computer conversations are relatively few. The Rapport agent [44] observes the head movements and voice prosody of a user telling a story, and generates contingent visual listener behaviour including nods and posture shifts. It produces no vocal feedback, and it never speaks. Its automatically generated listener behaviour was rated about as well as natural human listener behaviour in a face-to-face condition, and significantly better than a non-contingent version of the system. Using a variant of the Rapport agent, von der Pütten et al. [38] investigated the role of perceived agency and of behavioural realism on the feeling of social presence. Subjects rated mutual awareness more highly in the high behavioural realism condition; it did not matter whether the subjects believed they saw an autonomous ECA or a rendition of a human's behaviour. This seems to indicate that, given suitable expressive behaviour, it should be possible for an autonomous ECA to induce a sense of social presence.

The effectiveness of contingent emotional adaptation to a user's emotion in a spoken dialogue system was investigated by Acosta [45]. When the system adapted its prosody to the emotion detected from the user's speech, users rated the system as significantly better on a number of rapport scales, compared to a neutral baseline as well as a non-contingent version.

Adaptation also appears to occur in the other direction. Porzel and Baudis [46] investigated effects of a dialogue system on the user's non-verbal behaviour, and found a tendency for the human users to adapt to the non-verbal behaviour of their interlocutor. Human users produced significantly fewer vocal feedback signals, used less overlapping speech, and made longer pauses when interacting with a spoken dialogue system than when interacting with a human.

Despite these efforts, to the best of our knowledge, no full-scale dialogue system has been built before that takes into account the user's emotion and non-

verbal behaviour from visual and vocal cues, and interacts in real time both as a speaker and a listener in a multimodal conversational setting. It is also significant that none of the systems mentioned above is available as open source, which makes it difficult to improve existing work incrementally.

3 The Sensitive Artificial Listener scenario

When computational research tries to understand human abilities, one of the key challenges is to find portions of human behaviour that an artificial system has some chance of matching in a meaningful way. Marr [47] famously dismissed contemporary computer vision research on the world of blocks, and he was right. The 'blocks world' invited solutions that are mathematically elegant, but that have little or nothing to do with the way that human vision operates. It is widely accepted that the same holds for the worlds of grammatically perfect sentences, still photographs of posed emotions, and so on. On the other hand, it is counterproductive to insist that computational research is worthless unless it can unveil the subtle implications of blurred images where one party sneers in passing at the other's (off camera) shoes. Finding tasks that set appropriate challenges is one of the keys to progress.

The SAL scenario was invented explicitly with a view to setting a useful level of challenge. The idea was prompted by work with chat shows, where hosts appeared to follow a strategy that was simple and effective: register the guest's emotions, and throw back a phrase that gives very little, but that makes the guest more likely to disclose his or her own emotions. It seemed possible that machines could be programmed to carry out interactions of that general kind. They would need some rather limited kinds of competence, which there was a reasonable chance of achieving; but they would not need various other competences, which were much less likely to be automated in the foreseeable future. The main competences that (apparently) would be needed were recognising emotion from face, voice, and gesture; generating expressions that were emotionally coloured, but rather stereotyped; and managing basic aspects of conversation, such as turn-taking and backchanneling. Competences that (apparently) would not be needed included recognising words from fluent, emotionally coloured speech; registering the meaning and intention behind them; and generating a wide variety of emotionally coloured utterances and gestures 'from scratch', to meet the needs of the situation.

Several other situations reinforced the intuition that humans are capable of interactions that depend on sensitivity to emotion, but not much else. One is the kind of interchange that takes place at parties, where noise levels make it very difficult to understand the other party's words, but the emotional messages that are interchanged are often quite strong. Another is interaction between people who speak different languages, but who manage to interact at length by registering the emotional signs that the other party is giving. There is a prima facie case for thinking that modelling situations like these offers computational research an opportunity to develop a significant constellation of non-verbal capabilities without being distracted by difficult problems in speech recognition and natural language processing.

Translating that broad conception into a workable scenario raises a variety of issues, and they need to be solved in a coherent way. The SAL scenario is based on identifying the simplest solutions that create a degree of engagement. There has to be a representation that covers any affective states that the user is likely to display in this kind of interaction. Representations in terms of dimensions provide a straightforward way to do that. Specifically, the two most widely used dimensions of valence (negative-positive) and arousal (active-passive) (see e.g. [48]) are a natural starting point. There have to be rules relating the system's comments to user state. SAL uses the simplest option that produces interesting behaviour. At any given time, the system has a preferred state. If the user appears to be in that state, the system indicates approval; if not, it tries to propel him/her towards it. Systems with a single preferred state quickly become boring, and so a way of introducing different preferences is needed. SAL does it by creating a distinct character for each preference, which is much easier than moving a single character convincingly between affective states. Because user states are defined in terms of dimensions, so are the characters. The simplest option is to create one for each quadrant of valence-arousal spaces, and so that was done, with one qualification. Hence the system contains four distinct characters, each associated with an area in arousal-valence space. Spike is angry (negative-active), and tries to make the user equally angry; Poppy is effervescently happy (positive-active), and tries to make the user equally happy; Obadiah is despondent (negative-passive), and tries to make the user equally gloomy. The positive-passive quadrant remains empty because although we tried to devise a character to represent it, we did not find that it lent itself easily to sustained engagement. Instead we introduced a character who represents the emotionally neutral centre of arousal-valence space, Prudence, who is matter-of-fact, and tries to make the user equally matter-of fact. It is worth stressing that the point of these choices was to define a rational starting point for a long-term effort – not, for instance, because of a whimsical fascination with the characters. It is a natural goal to explore more complex choices as the technologies needed to deal with these ones are consolidated.

Note that the term Sensitive Artificial *Listeners* does not imply that the agents never speak. It reflects the fact that the flow of information is asymmetrical: the agents invite the user to disclose information, but do not disclose anything of substance themselves. Broadly speaking, that is the kind of behaviour that qualifies a person to be described as a good listener.

In some respects, the dialogue paradigm in the SAL scenario is a descendant of Weizenbaum's ELIZA [49]: both are systems that can draw listeners into extended interactions although they have no real understanding of what their words mean. But although they have that in common, SAL is diametrically opposite to ELIZA in fundamental ways. ELIZA relies on the massive simplification of input/output provided by a keyboard and on-screen text: these allow it to make use of (rudimentary) language processing skills, primarily textual pattern matching. In contrast, SAL has multimodal interaction capabilities – it can analyse the user's non-verbal behaviour through the analysis of voice and facial expression, and it will react through head movements, facial expression, and voice. The competences that let it use these depend very little on language. Keyword spotting is used, but the keywords are interpreted only in terms of their emotional connotation, not in terms of any domain concepts. Rudimentary understanding of conversational structure is also needed.

The SAL's utterances are drawn from a script of predefined phrases, which are used to introduce topics and to encourage the user to follow-up on a topic. Phrases are not used in a set order. What the system says at any given time is chosen according to two types of criteria. As indicated earlier, the selection depends on the user's emotional state at the time, and it is designed to attract the user to the SAL's own state. For example, with a negative passive user, Poppy would use sentences such as "There must be good things that you remember!", whereas Spike would rather say "Life's a war, you're either a winner or a loser." If the user is in the same state as the character, agreeing and reinforcing phrases are used such as "I love to hear about all this happiness." (Poppy) or "It wears you down, doesn't it?" (Obadiah). The second kind of selection criterion deals with conversational structure. That involves, for instance, using phrases like 'tell me more' when the user appears to be interested in a topic, and keeping repetition at an acceptable level.

In a SAL session, a user first receives an introductory briefing on the scenario and on the four characters and their limitations. He or she can then choose one of the characters to talk to first. After interacting with that character for a while, the user can switch to a different character, and thus can talk to all four characters within one session.

The SAL scenario has repeatedly been shown to work well with users who are willing to engage with the system (see Section 4); it is not designed to engage users who do not engage by themselves. It is very easy to break the system, e.g. by simply not talking. Because of this, the introductory briefing is essential before users interact with any version of the SAL system. Despite this limitation, the SAL system is generic in the sense that it allows a free dialogue with the user about anything, in real time.

4 HUMAN-DRIVEN SAL PROTOTYPES

The concept of SAL was outlined in the previous section. Translating it into a working system involved a series of prototypes in which the SAL's behaviour was determined, in one way or another, by a human operator.

These prototypes addressed two issues. One was checking whether the SAL scenario was indeed suited for sustaining a conversation based essentially on emotional and non-verbal information, and making adjustments where necessary. The other was acquiring the data needed to build the relevant system components. The two functions are logically separate, but practically intertwined: recordings of tests provide the data. This section first describes the prototypes and their logical functions, and then considers the data.

Table 1 points out key properties of the three prototype versions and the Autonomous SAL system described in Section 5.

4.1 PowerPoint SAL

In the first stable version of SAL, an operator responded to each user utterance by choosing an appropriate sentence from a range of options displayed on a monitor in front of him/her, and delivering it in a tone of voice that suited the SAL character who was then in play. The options available were dictated by the operator's judgment of the user's current emotional state, because each screen showed only utterances that the character then in play was deemed likely to make to a user in that state. A screen would typically present 20-30 options, and included buttons that the operator clicked to bring up a different screen if the user changed mood or asked to speak to a different character. The system was implemented in PowerPoint (hence the name). User and operator faced each other in the same room, but eye contact was minimal because the operator was usually concentrating on the screen.

The work with PowerPoint SAL confirmed that people can indeed engage in interactions of this kind for sustained periods [50]. Most users (though not all) responded with what raters judged was genuine emotion, often quite intense. The problems that arose allowed the scripts to be refined, and provided insight into the selection strategies that effective operators used.

TABLE 1: Key properties of the three SAL prototypes and of the autonomous SAL system.

	PowerPoint SAL	Semiautomatic SAL	Solid SAL	Autonomous SAL
SAL behaviour determined by...	operator	operator	operator	software
User sees...	operator (same room)	flickering bars	operator (on screen)	virtual character
User hears...	operator	pre-recorded audio	operator	synthetic speech
SAL utterances strictly follow script	yes	yes	no	yes
Setup allows for SAL to show listener behaviour	yes	no	yes	yes
Setup allows for high-quality recordings	no	yes	yes	yes

4.2 The SEMAINE test environment

The next steps depended on an environment which provided both high quality recording and full control over the information flowing between users and operators [10]. User and operator sat in separate rooms. Each looked into a tele-prompter, which consists of a semi-silvered screen at $45°$ to the vertical, with a horizontal computer screen below it, and a battery of cameras behind it. The operator's screen always showed the user's face: the user might see the face of a human operator or a synthetic display. The fact that the cameras are behind the teleprompter screen means that where the user saw the face of the human operator and vice versa, both parties could have the impression of looking directly at each other, including eye contact. The setup also included multiple microphones for each participant, routed through processors which allowed sound to be filtered or attenuated. A specially designed computer with multiple hard disc drives was needed to capture data from all these sources [10].

4.3 Semiautomatic SAL

The logical successor to PowerPoint SAL [51] was known as 'Semiautomatic SAL'. It used the SEMAINE test environment. Instead of reading the relevant utterances (as in PowerPoint SAL), the operator clicked on them, and a recorded version of the utterance, in a voice suited to the character, was played. The user heard the recorded utterance and, on the screen saw a schematic face, with coloured bars beneath it that flickered in synchrony with the speech. The point of the display was to give users a focus of visual attention: the synchrony between bars and voice creates a perceptual connection between them.

A key function of this system was to verify that the task being set for the autonomous system was achievable. Because of the limitations of automatic speech recognition, choice of utterances in the autonomous system would have to be based almost wholly on nonverbal cues – facial expression, vocal signals, and so on. In contrast, PowerPoint SAL operators had been able to use the user's words to guide their choice of response.

Studies with Semi-automatic SAL tested whether the verbal component was essential. Each participant interacted with all four characters. The operator had full sound in two of the cases, and it was degraded in the other two. After interacting with each character, users answered questions developed to assess the quality of the interaction. They are described in Section 6, but in broad terms they deal with three areas: perceived flow of conversation, inappropriate statements, and engagement. Two levels of auditory degradation were used. In the first, an acoustic filter allowed the operator to hear prosody, but very few identifiable words. Interaction was almost unimpaired, and so in the next stage the degraded condition provided the operator with no sound at all (though the user's face was still visible). User ratings were only slightly worse in this condition than with undegraded audio. This suggests that the SAL scenario meets one of its key goals: it is a type of interaction that can be conducted without understanding the user's speech, provided that the operator – human or automatic – can read the relevant non-verbal information sufficiently well.

4.4 Solid SAL

There are important kinds of behaviour that cannot be studied with Semiautomatic SAL, most obviously backchanneling behaviour by the SAL character while the user is speaking. That is clearly important for an agent's ability to keep a human engaged in an interaction, but it cannot be generated by the operator of a Semiautomatic SAL system, since he/she spends most of his/her time searching options on a screen and clicking on links.

A third scenario, called Solid SAL, was developed to provide the relevant data. The key feature of the Solid SAL format was that the operator did not read responses from a screen. Instead he/she was expected to be closely familiar with each of the SAL characters' properties, and tried to speak as the relevant character would do. The main function of Solid SAL was to generate data, and it is considered in the next section.

4.5 Data and annotation

The primary motive for recording the interactions was to provide data relevant to the design of various SEMAINE system components. For example, the Solid SAL scenario was devised to capture the kinds of user behaviour that were relevant to training the emotion recognition and feature detection systems. These recordings informed the visual and auditory recognition components. Similarly, the recordings of

TABLE 2: Amounts of recordings and annotations for each of the SAL scenarios.

	PowerPoint SAL	Semiautomatic SAL	Solid SAL	Autonomous SAL
Character Interactions	96	144	95	640
Unique Users	22	36	20	80
Duration (mins)	135	720	475	1920
Recordings show	User only	User only	User & operator	User & agent
Transcripts	Yes	No	Yes	No
Raters per clip	4	2	6-8	1
Dimensions per clip	2	9	9	1
Accessible from	www.psych.qub.ac.uk/PowerPointSAL	www.semaine-db.eu	www.semaine-db.eu	www.semaine-db.eu

the operator informed the synthesis of agent turn-taking and dialogue management. The same motive underlay the choice of labelling methods. They were designed to capture, what the system would need to know about the user in order to respond appropriately, in a format that was coherent and tractable. For that reason, the form of the labelling was shaped by extensive consultation between psychological and technological partners.

This section aims to summarise the main features of the labelled data in view of their relationship to system development and evaluation. A second motive for the recordings was to provide data for future research on fluent, emotionally coloured spoken interactions. A separate paper provides in-depth description of the data with that aim in view [52].

As explained in Section 3, the SAL scenario was specifically designed to take advantage of the generality and simplicity that dimensional representations provide, with particular emphasis on the two best known, valence and arousal. PowerPoint SAL recordings were annotated with only those two dimensions. For SEMAINE, it was judged that richer descriptions could be useful. Power and Anticipation/Expectation were chosen as additional dimensions because they were identified as key dimensions in an influential recent study [53]. Intensity (which is the longest established dimension of all) was also added. It is sometimes confused with arousal/activation, but "extremely intensive anger is likely to be characterized by high arousal whereas intense sadness may be accompanied by very low arousal" [54, p. 719]; and experiments confirm that they are psychologically distinct [55].

Trace-type descriptors were used to annotate recordings on these dimensions. A trace is a continuous-valued record of the way some aspect of the target individual's state appears (in the eyes of a rater) to rise and fall over time. Annotations of PowerPoint SAL used the FEELtrace tool, which provides two traces simultaneously [56]. Annotation of the SEMAINE recordings used the more recent type of trace tool [57], where each dimension is traced separately. Raters are presented with a scale running from the lowest possible value for the attribute in question to the highest, and they 'trace' the user's apparent state by moving the cursor up and down the scale as it rises and falls. In the form used for SEMAINE, the reliability of that technique compares well to the more obvious strategy of assigning everyday emotion words, and it is more suitable for the training needs of the SEMAINE system components [58], [59].

Many other types of description also seemed potentially useful, but it was not feasible to cover them all comprehensively. The solution adopted was to present tracers with a wide range of descriptors spanning 'basic emotions'; what Baron-Cohen [60] has called affective epistemic states (states which involve both knowledge and feeling); communicative categories; and descriptions of the quality of the interactions. From that list, they chose the four attributes that appeared to be most relevant to describing the extract in question. (Each extract was a single user's interaction with a single SAL character. These lasted about 3 minutes with Autonomous SAL, and on average slightly longer with the other versions). Specifying four attributes as the norm reflects the level of detail that labellers using earlier systems felt they could meaningfully give [57]. They then carried out a trace for each of those attributes (for instance, if they chose anger, they traced the apparent rise and fall of anger through the extract; if they chose 'agreeing/disagreeing', they traced the extent to which the user appeared to be displaying agreement; and so on). The result was that ratings of these categories were made when they appeared to capture a significant feature of the interaction, but not otherwise. Analysis of the data [52] vindicates the decision to limit the attention devoted to these categories. Most of the information is captured by a small number of descriptors – two thirds of the choices are accounted for by seven of the labels (happy, amusement, at ease/not at ease, thoughtful, expresses agreement/disagreement, gives information, gives opinion).

The full tracing system that has been described was applied to recordings of Solid SAL and Semiautomatic SAL. Time did not allow the same procedure to be applied to the Autonomous SAL recordings, but a different trace was applied to them in the course of evaluation. It rated the user's apparent level of engagement. Separate analyses were carried out to provided data on specific issues, including facial action units and head nods/shakes (see Section 5.2.2). Cowie et al. [61] have reported statistical analyses of the head movements involved in backchanneling during Solid SAL interactions.

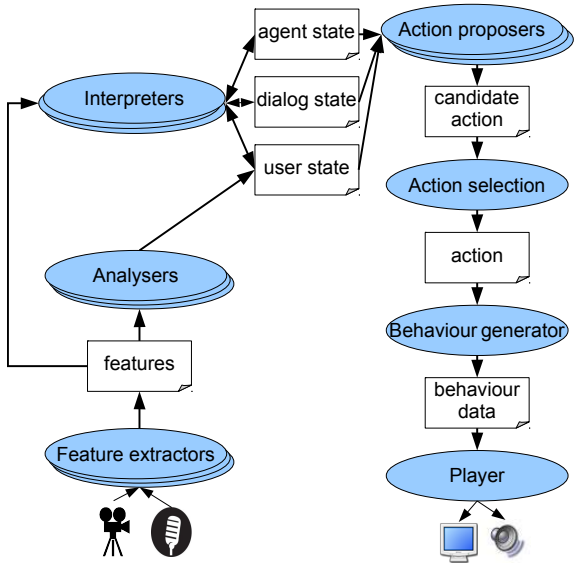


Fig. 1: Conceptual architecture of the SAL system

Table 2 summarises the available SAL material.

5 Autonomous SAL

Having motivated the SAL scenario and reported on prototypes controlled by human operators, we now describe our implementation of a fully autonomous SAL system. It is not controlled by any operator; all of the SAL characters' behaviour is controlled by the software components described in this section. We report on our approach to integrating the system, conceptually and technically, and describe our solution to the implementation of the different system components.

5.1 Building an integrated system

The integration of multiple input and output components into the real-time interactive SAL system is facilitated by a conceptual and a technical framework.

5.1.1 Conceptual framework

Conceptually, the architecture that orients the implementation of the Autonomous SAL system is very similar to the architectures of other multimodal interactive systems as described in Section 2. Fig. 1 shows the main items. User behaviour is observed through a camera and a microphone, and low-level features are computed using a battery of *feature extractor* components. A set of *analyser* components make individual analyses of the user's non-verbal behaviour and emotion, which are merged by *fusion* components which also produce an estimate of the information's reliability. *Interpreter* components process the fused analysis results in the context of all known information, and take decisions about the system's 'current best guess' about the state of the user, the dialogue and the agent.

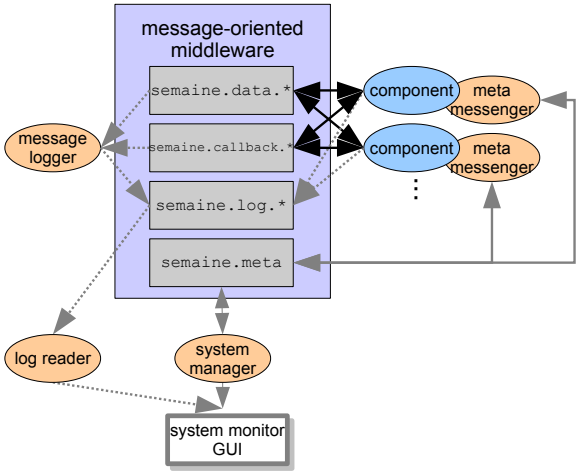


Fig. 2: Architecture of the SEMAINE API component integration framework

In parallel, a group of *action proposers* continuously take decisions on whether to propose an action given the current state information. An *action selection* component makes sure only one action is being realised at a time. The selected action is then prepared in terms of concrete vocal, facial and gestural behaviour, and finally rendered by a *player* component.

All components are described in some detail below.

5.1.2 Technical framework for component integration

We have created a custom, cross-platform component integration framework, the SEMAINE API [11]. Since our research system is built from components developed at different sites in different programming languages and operating systems, the framework is necessarily cross-platform and distributed: the components of a single integrated system can be spread over, e.g., three computers running Windows, Mac OS X and Linux, respectively.

The communication architecture is shown in Fig. 2. Components can send each other data messages, which transport information, as well as callback messages, which inform about processing states (e.g., player started/finished playing a certain utterance). Each component has its *meta messenger* which stays in contact with a *system manager* keeping track of the state of the overall system. The system manager can display a message flow graph in a *system monitor* window. By means of a centralised logging mechanism, the system monitor can also display log information from all components, as well as the messages that are being sent. This assists the developers in the task of keeping an overview of what is happening in the distributed system.

This architecture makes systems built on top of the SEMAINE API highly modular, and it is an explicit design goal to support reuse of components or subsystems. Therefore, the data sent between components use standard representation formats where possible.

For example, the Extensible Multimodal Annotation (EMMA) language [62] is used for representing the output of *analysers*; the Behaviour Markup Language (BML) [63] is used for representing the SAL agent's behaviour in the process of realising actions, etc.

5.2 Feature extraction

All understanding of the user's behaviour that the SAL system may achieve starts with the extraction of low-level features characterising the user's voice, head movements, and facial expression.

5.2.1 Acoustic features

As a first step, voice activity detection is applied in order to extract meaningful acoustic features only in regions where the user is talking. Two methods are considered in the SEMAINE system: a simple method based on a signal-energy threshold, to be used for quiet environments, and a self-adaptive voice classification method for noisy conditions.

A set of 1 882 acoustic features is extracted from the audio signal per analysis segment. Features are generated by applying statistical functionals to contours of acoustic low-level descriptors. Among the low-level descriptors are commonly used features such as loudness, fundamental frequency, probability of voicing, Mel-Frequency Cepstral Coefficients (MFCC), and other features based on the signal spectrum. The functionals include common statistical descriptors such as mean, standard deviation, and other analytical descriptors [64], [65].

For keyword detection we use a large vocabulary continuous speech recognition engine tuned for robust recognition of spontaneous and emotional speech [66]. In addition to conventional MFCC features, the system uses phoneme predictions generated by a context-sensitive neural network which was shown to be well-suited for modelling conversational speech [67], [68]. The keyword detector uses acoustic models and language models trained on the PowerPoint SAL and Solid SAL databases as well as the COSINE corpus [69] consisting of conversational, disfluent, and partly noisy speech. We trained additional acoustic models for non-linguistic vocalizations, such as *laughing*, *breathing*, and *sighing*.

In order to enable combined acoustic and linguistic emotion recognition, binary bag-of-words features are computed from the keyword output [12]. These features indicate for each word in the 'bag' whether it is present or absent in an analysis segment. The most relevant keywords differ from one emotion dimension to another. Approximately 200-300 keywords per dimension have been determined by a statistical feature selection method.

5.2.2 Visual features

The nonverbal visual events that we aim to recognise are head gestures and facial action units (AUs). Head gestures are characterised by the amount and direction of head motion, as well as occurrences of head nods/shakes and head tilts (right/left). Recognition of the aforementioned visual events is based on the low level features extracted.

The computation of visual features from the video signal starts with a face detector [70]. The algorithm first searches for a frontal face, and if that is not found, it searches for a profile face.

After the face is detected, an algorithm similar to the face detector is used to detect the left and right eyes. Using the locations of the centres of the eyes, we compute the *roll* of the face, which is useful to detect head tilts and to remove appearance changes due to tilted heads.

For head tilt detection based on the roll of the face, we average the results over a time window of 0.4 seconds. If the average angle value is greater than 0.1 radians, we conclude that a right-head-tilt occurred, and if it is smaller than -0.1 radians, a left-head-tilt is detected.

In order to determine the magnitude and the direction of the two-dimensional head motion, the optical flow between two consecutive frames is computed. It is applied to a refined face region within the area returned by the face detector to ensure that the target region does not contain any background information. The resulting optical flow vector represents the average horizontal and vertical head movement over all pixels in the face region. To simplify this representation and make head action detection more robust, motion is discretised into five groups that we call *directional codewords*: rightward, upward, leftward, and downward movement, and no movement.

A so-called appearance-based representation of the face is important for various tasks, such as facial expression recognition and face recognition. In our system we describe the appearance of the face using Local Binary Patterns (LBPs) [71]. For any image, a histogram of the LBP output describes the texture of that image.

To describe local texture instead of a single texture for the entire face, we divide the face region into $10*10$ blocks. An LBP histogram is calculated for each of those blocks separately, after which the histograms of all blocks are concatenated to form a single feature vector. To be able to deal with appearance variation (in terms of head roll and face size), we use the locations of the eyes found during head tilt detection. The input images are first rotated to neutralise the face *roll*, and then scaled to a uniform size.

Facial Action Units (AUs) are detected based on the appearance-based facial features described above. After a feature selection step, the reduced feature set is fed to a bank of Support Vector Machine classifiers, one for every AU detected. The system can reliably detect AU 1 (moving the inner eyebrows up), AU 2 (moving the outer eyebrows up), AU 4 (moving the

eyebrows down, i.e. frown), AU 12 (lip corners pulled up, i.e. smile), AU 25 (lips parted) and AU 45 (blink) [65]. These action units are used in the generation of mimicry-type listener behaviour (see Section 5.4).

5.3 Understanding human behaviour

This section describes the algorithms and methods used to analyse human behaviour based on the acoustic and visual features.

5.3.1 Emotion analysis from acoustic features

The Autonomous SAL system contains speech-based detectors for five emotion dimensions: arousal, valence, expectation, intensity, and power on a continuous scale from -1 to +1. The acoustic feature vector is concatenated with the binary bag-of-words vector for the respective emotion dimension, resulting in five different feature vectors for the respective five dimensions. We use Support Vector Regression [72] to map the feature vectors to dimensional affect analyses.

For experimentation and evaluation we chose recordings from the Solid SAL database [10] that have been annotated by multiple raters. Adding the linguistic features (keywords from the real automatic speech recogniser output, *not* from the ground truth transcription), improves the results compared to a purely acoustic feature set. Our results indicate that in particular the valence dimension is difficult to detect reliably from acoustic features alone; automatically detected keywords, despite their limited accuracy, improve the results substantially. Detailed results are reported in [65].

Additionally, an interest detector is included as an attempt to capture the user's interest in the conversation, which uses a Support Vector Machine classifier to identify three discrete classes of the level of interest. The classifier was trained on the TUM Audio-Visual Interest Corpus (AVIC), which was recorded explicitly for the purpose of automatic identification of the user's level of interest [73]. Since some components in the SEMAINE system require a continuous *level of interest* value, a continuous estimate of user interest is computed as the centroid of the three discrete labels, by weighting each label with the confidence probability returned by the classifier.

5.3.2 Emotion analysis from nonverbal visual events

Emotion analysis is based on the detection of head nods and shakes.

For head nod and shake detection, training data were obtained by manually annotating examples of nods and shakes in the Solid SAL [10] database. The directional codewords generated as part of the visual feature extraction module were fed into a Hidden Markov Model (HMM) for training a nod model and a shake model. In order to analyse the visual data continuously, we empirically chose a window size of 0.4 secs that allows the detection of both brief and longer instances of head nods and shakes [13], [14]. We use a number of measures in order to be able to distinguish other head movements from the actual head nods and shakes: (i) we threshold the magnitude of the head motion; (ii) we built an 'other' model to be able to recognise any movement other than nods and shakes; and (iii) we statistically analyse the likelihoods outputted by the nod/shake/other models.

Emotions are detected from results of the head gesture analysis, i.e., the head motion features and the detected nods/shakes. We detect the same five dimensions as for speech: arousal, valence, expectation, intensity and power. The system uses multiple detectors trained by using the ratings from each human rater and each dimension separately. The labels used for training the automatic detectors (i.e., ground-truth) for a given window consist of the dimensional emotion trace annotations averaged over that window. Such a representation allows us to consider each feature vector independently of the others. We used Support Vector Regression for dimensional emotion detection as this technique has proven to be efficient for various regression tasks [74], [75].

TABLE 3: Pearson's correlation coefficient (CC) for the fused emotion analysis averaged over three raters, for five emotion dimensions: (A)rousal, (E)xpectation, (I)ntensity, (P)ower, and (V)alence. The average correlation between the other two human raters and the respective rater is provided as a reference.

Dimension	A	E	I	P	V
CC	0.28	0.18	0.36	0.23	0.24
CC (human)	0.51	0.37	0.40	0.35	0.51

5.3.3 Fusion-based emotion analysis

The fusion procedure in the Semaine system is a very simple one: the fused emotion value for each dimension is computed as the weighted average of the values estimated for that dimension by each of the individual classifiers. The confidence value produced by each classifier is used as the weight for that classifier's contribution to the fused value.

An evaluation of the fusion procedure was performed by correlating the fused emotion value for each dimension with human ratings. Since separate classifiers were trained for each human rater, the fused values were correlated with test data from the same rater (not used in training). As a reference, the average correlation between *other* human raters and the given rater is also computed. Table 3 presents both classifier and human correlations on two test recordings, averaged over three raters.

Overall, our results show that obtaining a high correlation between various human raters, for the audio-visual Solid SAL data, is indeed challenging. To mitigate a similar problem, [76] proposed a method

for achieving high inter-rater agreement and segmenting the continuous sequences into shorter clips. However, extending this method for automatic, dimensional and continuous emotion detection remains a challenge. Moreover, during annotation, the human raters were exposed to audio-visual Solid SAL data without explicitly being instructed to pay attention to any particular cue (e.g., head gestures) or modality (e.g., audio or video). Therefore, it is not possible to conclude which cues or modalities were dominant for which ratings. In general, these issues still remain as open research questions in the field [74], [75].

5.4 Dialogue management

In this section, we describe how *interpreters* and *action proposers* work together to interpret the user's behaviour and to determine the agent's *turn taking* behaviour (*when* the agent speaks), its *utterance selection* (what it says), and its *backchanneling* behaviour while it is in the listener role.

The main challenge of these components is to interpret what is known of the user's behaviour and to generate natural reactions, without knowing much about the content since linguistic analysis is intentionally limited to crude keyword spotting in the SAL scenario. From the feature extractor and analyser components, the dialogue manager receives low-level features such as the energy of the audio, the fundamental frequency, the position of the detected face, and the direction of head movement. Analysers provide higher level features such as the arousal and interest of the user, facial expressions, and head gestures. Out of these, the emotional state of the user is the most important information.

5.4.1 Speaking SAL

As was explained in Section 3, the dialogues with the SAL characters are rather special from a computational point of view. Basically, the agents are chatbots that do not attempt to understand what the human interlocutor is saying and that do not have a very defined task they want to see performed. As chatbots go, what the human interlocutor can say is left open and uncontrolled, whereas the SAL characters each have a limited repertoire of canned phrases they can choose from. This means that the role of the dialogue manager in the SAL system is to pick out the most appropriate sentence to say at any given time. The adequacy of the choice of sentence is determined by two main criteria. The basic one is whether the agent keeps the interlocutor involved in the conversation: *sustained interaction*. For this reason, many SAL utterances are prompting the user to say more. The second criterion is determined by the SAL 'goal' to draw the user towards the character's emotional state.

The dialogue manager consists of a number of utterance selection modules that each focus on a particular criterion for selection. For example, an 'After Silence' module suggests responses to occur after a long period of user silence. It includes responses such as "Well?", and "Go on, tell me your news!". These responses are used to motivate the users to continue speaking if they are silent. A 'Linking Sentence' module suggests a follow-up response to a previous question of the agent depending on the answer from the user. For example, when the agent asks "Have you done anything interesting lately?", and the user responds with a short answer with an agreement in it, a linking sentence could be "You did? Great! Please tell me about it.". A 'Content' module suggests responses based on the detected keywords for a number of high-level categories such as 'talking about past', 'talk about own feelings', 'agree/disagree', etc. When keywords in the user's utterance match one of these categories, this module proposes responses from the respective category set. An 'Audio Feature' module suggests responses based on the detected audio features of the user's turn. Classifiers that have been trained on the Solid SAL data complemented with human suggestions of good responses provide the suggestions for SAL [77]. An 'Arousal' module suggests responses based on either a very high or a very low arousal of the user. For example, Obadiah might say "Don't get too excited" after detecting high arousal, and Prudence might say "You seem a bit flat" after detecting low arousal. Finally, a 'Backup Responses' module suggest some generic responses that fit in most of the cases. This includes responses such as "Really?" and "Where do you think it will lead?".

Each module returns a list of possible responses – possibly empty – with an estimate of the quality of each response. The quality of responses is lowered for those responses that have been used in recent turns. The response with the highest value is selected and sent to the *action selection* component (see below).

5.4.2 Listening SAL

A separate *action proposer* component generates the agent's listener behaviour. From the literature [78], [79] we have fixed some probabilistic rules to decide *when* a backchannel signal should be triggered. Our system analyses the user's behaviours, looking for those that could prompt an agent's signal; for example, a head nod or a variation in the pitch of the user's voice will trigger a backchannel with a certain probability. As the next step, the system calculates *which* backchannel should be displayed. The agent can provide either *response* signals that transmit information about its communicative functions (such as agreement, liking, believing, being interested and so on) [3], [80] or signals of mimicry that mirror the speaker's non-verbal behaviour directly.

The *action selection* component [81] receives all the candidate actions coming from the *action proposers*. Actions vary in complexity, ranging from a single head

nod, via multimodal backchannel signals involving, e.g., a combination of smile, nod, and "uh-huh", to full verbal utterances with synchronous non-verbal expressivity.

The action selection has two roles. The first one is to manage the flow of candidate actions to be displayed by the agent. Indeed the *action selection* continuously receives candidate actions. These are queued, since only one action can be displayed at a time. The action selection waits until the display of the current action has been completed before selecting another one. Speaker actions are given a higher priority than listener actions in the selection.

In the listener mode, the second role of the action selection is to choose the most appropriate backchannels to be displayed. This selection varies with the four personalities of the SAL characters [17], [82]. It also takes into account the emotions and interest level of the user.

Fig. 3: The four SAL characters as they appear in Autonomous SAL: aggressive Spike; cheerful Poppy; gloomy Obadiah; and pragmatic Prudence.

5.5 Generating SAL behaviour

Once the dialogue components have determined whether the agent is in a speaking or a listening role, and how it should act given that role, its behaviour must be realised. We use the same components for generating both speaking and listening behaviour.

The *behaviour generator* component receives as input a representation of communicative functions to realise for the current SAL character. Its task consists in generating a list of behavioural signals for each communicative function; this process draws upon a definition of behavioural characteristics for each of the SAL characters, which we call that character's *baseline*. An agent's *baseline* contains information on that agent's preference for using a given modality (speech, head, gaze, face, gesture, and torso) [83]. For the visual modalities, the baseline specifies also the expressive quality. Expressivity is defined by a set of parameters that affect the qualities of the agent's behaviour production, such as wide vs. narrow gestures, or fast vs. slow movements. A behaviour *lexicon* associates communicative functions with the corresponding multimodal signals that an agent can produce. Depending on the agent's baseline and the communicative function to convey, the system selects in the lexicon the most appropriate multimodal behavioural set to display. For example, an agent that wants to communicate its agreement while listening could simply nod, or nod and smile, or nod, smile, and say "m-hm".

The *audio synthesis* component [84] synthesises both the spoken utterances and vocal backchannels like *myeah*, *uh-huh*, *oh*, etc. For the spoken utterances, we created expressive unit selection text-to-speech voices [85]. The system also generates vocal backchannels [86]. In order to achieve lip synchronisation for verbal utterances and for audiovisual backchannels, our implementation generates timing information together with the speech, using the same timing representation formats for text-to-speech and for listener vocalisations.

Finally, the multimodal behavioural signals are transformed into animation parameters. Facial expressions, gaze, gestures and torso movements are described symbolically in repository files. Temporal information about the speech and listener vocalisations, generated by the *audio synthesis* component, is used to compute and synchronise lips movements.

The animation is played using a 3D graphics player showing one of the four SAL characters at a time [87].

Custom facial models for the four 'Sensitive Artificial Listeners' were created by a graphics artist (see Fig. 3). To create the models we collected information from the literature on personality traits. We also gathered data through illustrations, photos, video corpora, etc. All these materials helped us to specify the design parameters for our characters [17], [82].

6 Principles for Evaluating Sensitive Artificial Listeners

It has been clearly recognised for some time that evaluating systems concerned with emotion and affect presents particular challenges [88]. Evaluating the SAL system brings together several of the challenges. This section deals with the principles of evaluation before describing an evaluation of the Autonomous SAL system.

The system calls for evaluation at several different levels. As a first approximation, lower level issues can be separated out and addressed in comparatively straightforward ways — for instance, by measuring how often emotion is identified correctly from voice alone.

6.1 Principles for low-level evaluation

Low-level evaluations of the various components have been described above (see Section 5). Even those raise questions that are far from trivial, for reasons that have gradually become clear. The literature on speech provides a well-developed illustration. The obvious measure, percentage correct identification, was used as a metric in early studies. Collating their findings shows how inappropriate that is [89]: scores depend massively on both the number of classes being considered and the naturalness of the material. Providing a satisfactory alternative is not easy, but there has been interesting work on it.

Broadly speaking, the issues are linked to various kinds of distinctiveness that are inherent in the task.

First, it is natural to assume that success equals matching an ideal observer, but there are both general and specific reasons to question that. On a general level, it should no longer be in doubt that people differ in their perception of emotion-related material. One of the few extended descriptions, by Cowie and Douglas-Cowie [90], indicates that individual raters weight emotion-related features of speech differently. Matching a single observer who is not eccentric, or (very much the same thing) the average of a group of raters who perform similarly, may be a more rational aim than matching an ideal or average observer. The particular context of SAL underlines the point. We might feel that Spike should pick up marginal signs of aggression where Poppy would not, and that is in line with evidence that mood affects the perception of emotion-related stimuli [91].

A second issue is that very different formats can be used to describe raters' impressions, and they invite different metrics. SAL raters provide continuous traces. Other groups favour categorical descriptions, in some cases using a small number of categories (positive/negative/neutral), in others starting with dozens of options. One way to achieve comparability is to reduce multiple labels to a few 'cover classes', and to reduce the traces to a few qualitative labels, such as positive, negative or neutral valence [92]. But while that kind of description may facilitate comparison, it is not necessarily what a working system like SAL needs.

Nevertheless, a third issue makes it very desirable indeed to establish some kind of cross-system comparison. SAL data is in some respects quite challenging, and recognition rates are not likely to be high. Hence it is essential to know whether observed rates are due to poor systems or difficult material. A very useful comparison is provided by a recent report of recognition rates on SAL and other corpora using standard speech technologies [92]: the reported rate is 57.8% correct for a binary decision, lower than the standard AIBO corpus (62.9%) but higher than Smartkom (53.9%). Building up a broadly based understanding of different databases, and the kinds of recognition rates that they support, is a complex task [93]; but there seems to be no alternative way of gauging what particular scores on a particular database mean.

Beyond all that, it is not necessarily the case that the module which scores best as a stand-alone component is the most useful within a larger system. For example, it is notoriously hard to recognise valence from speech alone [94]. Hence while a purely speech-based module might improve its ability to recognise emotion classes by incorporating sensitivity to valence, the unreliable valence information that it used might actually degrade performance in a system that had access to much better valence information from vision.

6.2 Principles for high-level evaluation

Evaluating the system as a whole is not a routine exercise, and it raises questions which are of interest in their own right. Standard approaches are underpinned by the conception of usability stated in ISO 9241: the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." [95, p. 2]. Satisfaction has an affective component, but even it is defined in functional terms, as lack of discomfort, and a positive attitude towards the system, while performing the goals. The widely known evaluation framework PARADISE [96] is useful as a concrete illustration. It quantitatively assesses the quality of task-oriented dialogues in terms of usability measures. The objective of maximising user satisfaction is operationalised in terms of task success, and dialogue costs which quantify the efficiency of the dialogue at carrying out the task.

These criteria are not applicable to SAL-type dialogues, where no particular task is to be achieved. The issue is not specific to SAL: It covers a wide range of situations where "effectiveness, efficiency and satisfaction" are not the goals. As Edwardson put it, "We don't ski to be satisfied, we want exhilaration" [97, p. 2]. Growing awareness of that issue has led to the development of measurement systems that deal more directly with affective response. Work has addressed four main areas: computer anxiety; frustration; trust and loyalty; and 'flow, fun, and playfulness' [88]. The first appears not to be relevant for SAL. Frustration clearly is, and one might assume it was simply undesirable. However, that is not necessarily so. There is a kind of frustration that is a mark of human engagement. If we treat the SAL characters as people, then it is right and proper that we should be frustrated by Obadiah's relentless pessimism or Poppy's relentless brightness. Trust raises similar issues: a convincing Spike would not engender trust. The last area is partly similar. If Spike is convincing, an encounter with him is neither fun nor playful. However, it does create the characteristic 'flow' feeling of being engrossed in a task, to the exclusion of distractions [98].

Eliminating inappropriate questions exposes an issue that affects not only SAL, but any system designed to achieve conversational interaction. It is whether the system allows users to feel engaged in a fluent conversation with a credible personality; and linked to that, whether it draws them into the kind of behaviour that typifies conversational interaction with a person. These are closely related to the issues that have been highlighted in research on presence (e.g. [99]). The key challenge for the evaluation was to develop techniques that addressed those issues. Several options were explored (using Semiautomatic SAL and initial versions of autonomous SAL), and three were retained.

1) **Verbal probes.** Verbal reports are the obvious source of information, but there are well-rehearsed reasons for being wary of them in the evaluation of affective devices (e.g. [100]). Asking for verbal reports during an emotional experience, particularly one which involves suspension of disbelief, is likely to disrupt it; reports given afterwards are likely to rationalise it. The solution developed for SAL was, in effect, a spoken questionnaire designed to let users respond from within the scenario, with minimal disruption. Immediately following each interaction, a different avatar steps in and asks (orally) three questions about the interaction that has just finished:
   a) How naturally do you feel the conversation flowed? Zero means not at all, ten means totally natural.
   b) Did you feel the Avatar said things completely out of place? If yes, how often? Never, a few times, quite often, most of the time, all the time.
   c) How much did you feel you were involved in the conversation? Zero means not at all, ten means completely involved.

   The logic of the questions is that they deal with the three key entities involved: the avatar (specifically the appropriateness of its contributions); the user (specifically his/her feeling of engagement); and interaction between them (specifically its fluency).
2) **Behavioural engagement.** The study took advantage of a unique opportunity to obtain a trace-like rating of how engaged or disengaged users appeared to be. Because recording used two connected rooms, an observer could sit in the same experimental space as the user while the experiment was taking place, hearing the voice directly through the open door between the rooms and seeing him/her through the teleprompter used by the solid SAL operator. That gives a sense of the user's engagement or lack of it which, as the literature on presence would suggest, seems more compelling than observing a recording. A variant of the trace techniques used to label data was developed to capitalise on that. One of the raters whose scores correlated highly with group means in the labelling exercise ($r = 0.86$ for valence ratings) was recruited to observe the interaction as it took place, and to record the user's apparent engagement as it rose and fell with time using a mouse-controlled cursor.
3) **Simple preference.** The design of the experiments meant that users experienced two versions of the system. They were asked at the end which they preferred.

These measures provide a reasonable starting point for research on systems' ability to engage people in fluent interactions. They were used to evaluate the SAL system.

6.3 Evaluation of the final system

Evaluation involved a series of experiments. The earlier experiments highlighted problems with initial versions of the system [101]. Here we report only the final experiment. It does not claim to be a comprehensive evaluation. The point is simply to establish that the system's emotion-related abilities make a difference to the quality of human interactions with it. To do that, the final version of the system was compared with a control system in which affective features of the output had been disabled (the agents' voices and faces were expressionless, and they did not backchannel).

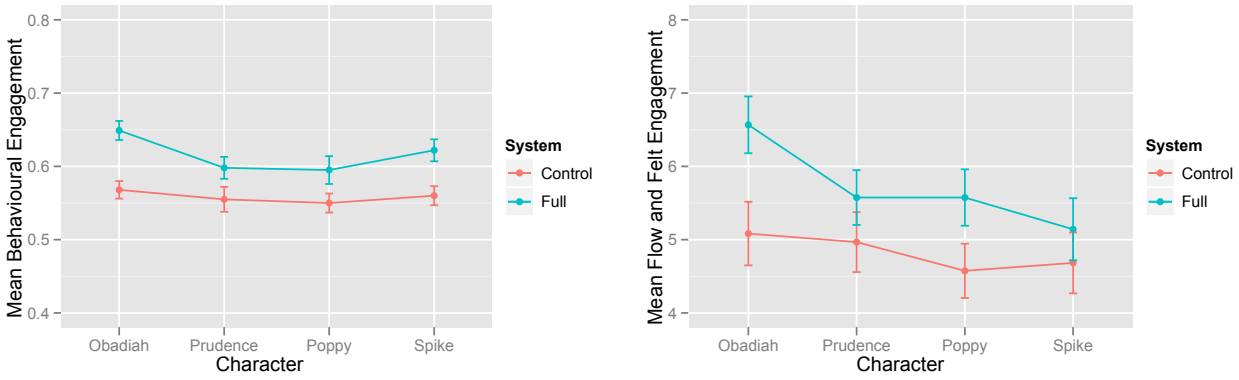
Participants: 30 users participated in the evaluation, 24 female and 6 male, ranging in age from 18-41 years old (M = 23.2, SD = 7.9) and all were from Britain and Ireland except one Polish and one Malaysian participant (both fluent in English). They were drawn from a pool of psychology students and received course credit for their participation. All participants were naive and had no or minimal experience interacting with a automated computer based avatar.

Design: A 2x4 within-subject design was used. System version was the first independent variable with two levels full system and control system. The four characters made up the levels of the second independent variable. The user interacted with all four characters twice, first with one version of the system, then with the other. The orders in which they encountered systems was counterbalanced and the order in which they encountered characters was randomised. Measurements were as outlined in section 6.2.

Procedure: The participants interacted with SAL avatars via the teleprompter arrangement used in all the SEMAINE recordings. At the beginning of each session, a non-affective avatar gave instructions for the interaction. After each interaction, the non-affective avatar presented the verbal probes (in the order flow, appropriateness and felt engagement). The

TABLE 4: Interactions that are predominantly in the upper 1/nth of the behavioural engagement scale.

Region within which most of the interaction lies	% of interactions that meet the criterion, with:	
	the full autonomous system	the inexpressive control system
Upper half	93%	86%
Upper 1/3 ("quite engaged" or better)	34%	6%
Upper 1/4	14%	0%
Upper 1/5	3%	0%



(a) Mean Behavioural Engagement (scale extends from 0-1)

(b) Mean Flow and Felt Engagement (scale extends from 0-10)

Fig. 4: Evaluation measures for the two systems across the four SAL characters (Error bars represent standard errors).

TABLE 5: Correlations between the four evaluation measures.

	Felt Engagement	Appropriate Contribution	Behavioural Engagement
Flow	0.82	0.48	0.48
Felt Engagement	—	0.54	0.42
Appropriate Contribution	—	—	0.27

Note: All correlations are significant at the $p < .001$ level.

behavioural engagement rater was following the interaction in a separate room adjacent to the recording room. The rater could observe the user's face on a monitor and hear the interaction through the door. The rater did not see the avatar.

A first question is whether the measures provide genuinely distinct kinds of information. Table 5 shows the correlations between the measurements. Flow and felt engagement correlated highly with each other, but they were not equivalent: engagement ratings were very consistently higher than flow ratings:in a MANOVA $F(1, 29) = 42.23, p < .001$. This is to say that what people feel the system does well is to engage them. Appropriateness of avatar contributions correlated moderately with the measures of both felt engagement and flow. The mean of the behavioural engagement measure correlated at a similar level with both subjective engagement and flow, and weakly with appropriateness of avatar contributions.

Analyses of variance show that the full system outperforms the inexpressive control on all of the measures. The main effect of system is significant for mean behavioural engagement ($F(1, 29) = 23.3, p < 0.001$); appropriateness of avatar contributions ($F(1, 29) = 4.86, p = 0.036$); and flow and felt engagement ($F(1, 29) = 9.25, p = 0.005$: these were treated as two measures in a single analysis because the underlying evaluations are closely related). The relevant means can be seen for behavioural engagement in Figure 4 (a) and for flow and felt engagement in Figure 4 (b). These give a partial picture, which is supplemented by other measures.

Simple preferences confirm that, as one might assume from the inferential statistics, the full system has a steady advantage: 24 out of 29 users who expressed a preference preferred the full system ($p < 0.002$ on a sign test). The behavioural engagement ratings provide a way of showing that the difference goes beyond a small but steady preference. They provide both information about timing, and a scale that has qualitative descriptors attached. The negative end is marked "absolutely no sense of engagement"; the boundary between bottom and middle third is marked "weakly engaged"; the boundary between middle and upper third is marked "quite engaged"; and the upper end is marked "compelling sense of engagement". That structure makes it natural to consider how many of the interactions stay in the upper 1/nth of the scale for most of the time. The figures are shown in Table 4. They capture what seems a reasonable assessment: sustaining a good level of engagement with the user is generally beyond the inexpressive system, but the expressive one achieves it quite regularly. Clearly it is very interesting to ask how the likelihood of that kind of interaction can be increased.

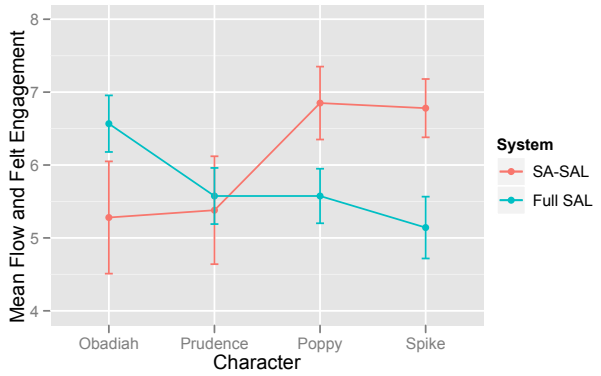


Fig. 5: Average of flow and felt engagement ratings for Semi-automatic SAL with full feedback to the operator, and the full version of Autonomous SAL (Error bars represent standard errors).

A second way of conveying how autonomous SAL performs is comparison with semi-automatic SAL, where a human operator chose what the system should say at any given time. Figure 5 compares the average of flow and felt engagement ratings for the full version of autonomous SAL with ratings for all the sessions where the operator had full feedback in Semiautomatic SAL. The key point is that the systems show very comparable ranges. Some characters fare better in one system, some in the other; but overall, the autonomous system is not grossly worse than one operated by a human.

A feature of the data is that they vindicate SEMAINE's decision to consider different characters. The effect of character is significant (with $p < 0.01$) in all of the analyses of variance reported above. The consistent finding is that Obadiah, the sad character, is rated best in the autonomous system. Prudence and Poppy follow, generally in that order. In contrast, the position of Spike, the angry character, varies considerably with both system and measure. Figure 5 makes the point that this is not a simple matter of poor scripts for Poppy and Spike: essentially the same scripts were well received in the semi-automatic system. These findings do not constitute a systematic study of the effect of agent character. They do indicate that agent character needs to be recognised as an issue in achieving emotional engagement: what works for an agent with one affective style may not work for another.

6.4 Conclusions and directions in evaluation

Two key points can be made with confidence. One is that the methods described above provide ways of measuring what distinguishes systems like SAL; the other is that the expressive abilities incorporated in the SAL system make a substantial difference to the interaction that a person can have with an autonomous system. It is equally clear, though, that evaluating systems of this complexity is a large task, which goes far beyond those points.

Evaluative data classically have two functions, formative (directing development) and summative (passing judgment). There is a great deal of scope for formative evaluation – identifying details that could improve the system. Some are reasonably clear, such as the timing of turn-taking; others are puzzling, such as apparent preferences for what were thought to be less competent versions of some characters. The major summative issue is which parts of the SAL system contribute. A natural way of doing that is to disable components one by one. That is technically difficult because the system was not designed with that kind of piecemeal adjustment in mind, and that issue is worth highlighting for future development.

Both should profit from additional evaluative techniques that have been piloted. Such techniques should allow users to signal dissatisfaction with minimal disruption to the interaction. Pilot work has also been done on objective analysis of behavioural engagement, by identifying gestures associated with low rated engagement [102]. Developed versions would also offer timing information, and, not least, enable future systems to evaluate their own success, by registering when the user was disengaging, and to take remedial action.

7 Conclusion

In this paper, we have presented a full implementation of an Autonomous Sensitive Artificial Listener. We have introduced the SAL scenario as a promising framework for investigating computational competence in dealing with emotional and non-verbal aspects of a conversation. We have described three prototype versions which were used to verify the postulated properties of the SAL scenario and to collect data needed for preparing the system components.

The Autonomous SAL system presented here is an integrated but highly modular piece of technology. We have shown one possible way of organising the structure of a SAL system, in terms of component architecture, message flow, representations of information, and processing steps. We have identified information that can be automatically deduced from the user's non-verbal behaviour with some degree of accuracy, and we have proposed a format for representing this information in terms of standard representation formats. We have realised a mechanism for generating both speaker and listener behaviour for ECAs exhibiting different personalities, again using standard representation formats in the workflow wherever possible. We have provided an implementation of the dialogue flow in the SAL scenario, and have provided a mechanism for flexibly extending or replacing the domain-specific information within the system.

We have laid out what we think are key principles to keep in mind when attempting to evaluate systems

such as SAL, and have shown that the expressive capabilities of the system improved the user's engagement over a non-expressive control system.

To the best of our knowledge, the Autonomous SAL system presented in this paper is unique in the current Affective Computing community, in that it covers the full loop of multimodal analysis, interpretation and synthesis of emotion-related and non-verbal behaviour in a human-to-computer dialogue setting, while at the same time being publicly available. In fact, the full system is available to the research community, in large parts as open source software, from the SEMAINE project website (http://www.semaine-project.eu). Interested researchers can download the system and adapt it at will to suit their interests. Many of the system's configuration options are accessible such that, even without programming, it is possible to create controlled variants of the system and carry out experiments. In the long run, the present system is suited as a baseline against which improvements of individual components can be assessed. Furthermore, the modularity of the system makes it possible to re-use individual components and build new, different emotion-oriented systems on the same platform and from existing and new building blocks. The use of standard representation formats is intended to promote and facilitate this process. We believe that in this way, the SAL system as presented here can have a lasting impact on the research landscape of interactive, emotion-oriented systems.

A number of scenarios lend themselves to exploiting and developing further the capabilities explored in the Autonomous SAL system. Relatively straightforward is the transfer to scenarios where the purpose of the interaction is a social or emotional one. An example is the scenario of a long-term interaction between a robot and hospitalised children [103], aiming to reduce their anxiety. This type of scenario lends itself to further exploration of SAL-type capabilities by means of interactions designed to concentrate on non-verbal and emotion-related aspects similar to the SAL scenario. A second short-to-mid-term scenario is that of multimodal games including social interaction, where the purpose of the emotional and non-verbal skills is to contribute to an overall immersive experience and high precision and recall are secondary. A more challenging step will be learning how to combine SAL-type capabilities with task-oriented interaction. Here, the purpose of emotion-related and non-verbal capabilities will be to strengthen engagement in the interaction, and to address barriers to effective interaction that arise from the user's emotional reactions. The challenge goes beyond adding expressive colouring to traditional task-driven dialogue: that would probably lead either to unnatural expressive behaviour or to perceived incongruence between the expressive capabilities and the task scenario. Understanding how to exchange information in the context of a fluent, emotionally coloured interaction will require extensive further research. The availability of the SAL system gives it a strong base from which to start.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE). We would like to thank our reviewers for their comments on previous draft versions of the paper.

REFERENCES

[1] R. Cowie, "Describing the forms of emotional colouring that pervade everyday life," in *The Oxford Handbook of Philosophy of Emotion*, P. Goldie, Ed. Oxford University Press, 2010, pp. 63–94, doi:10.1093/oxfordhb/9780199235018.003.0004.

[2] V. H. Yngve, "On getting a word in edgewise," in *Chicago Linguistic Society. Papers from the 6th regional meeting*, vol. 6, 1970, pp. 567–577.

[3] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, no. 1, pp. 1–26, 1992. [Online]. Available: http://jos.oxfordjournals.org/cgi/content/abstract/9/1/1

[4] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, p. 696–735, 1974.

[5] D. Jurafsky, J. H. Martin, and A. Kehler, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. MIT Press, 2000.

[6] J. F. Allen, D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, "Toward conversational human-computer interaction," *AI magazine*, vol. 22, no. 4, pp. 27–37, 2001.

[7] G. J. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N. Hawes, "Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction," in *Proc. Language and Robots*, 2007, p. 55–64.

[8] A. Mehrabian and S. R. Ferris, "Inference of attitudes from nonverbal communication in two channels," *Journal of Consulting Psychology*, vol. 31, no. 3, pp. 248–252, 1967. [Online]. Available: http://dx.doi.org/10.1037/h0024648

[9] R. L. Birdwhistell, *Kinesics and context*. Philadelphia, USA: University of Pennsylvania press, 1970.

[10] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, 2011.

[11] M. Schröder, "The SEMAINE API: towards a standards-based framework for building emotion-oriented systems," *Advances in Human-Computer Interaction*, vol. 2010, no. 319406, 2010. [Online]. Available: http://dx.doi.org/10.1155/2010/319406

[12] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867–881, 2010.

[13] F. Eyben, M. Wöllmer, M. F. Valstar, H. Gunes, B. Schuller, and M. Pantic, "String-based audiovisual fusion of behavioural events for the assessment of dimensional affect," in *Proc. IEEE Intl. Conf. Automatic Face and Gesture Recognition*, Santa Barbara, CA, USA, 2011, pp. 322–329.

[14] H. Gunes and M. Pantic, "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners," in *Proc. Intelligent Virtual Agents*, Philadelphia, USA, 2010, p. 371–377.

[15] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. Face and Gesture recognition*, Santa Barbara, CA, USA, 2011, pp. 314–321.

[16] M. ter Maat, K. P. Truong, and D. Heylen, "How Turn-Taking strategies influence users' impressions of an agent," in *Proc. Intelligent Virtual Agents*, Philadelphia, USA, 2010, pp. 441–453. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15892-6_48

[17] E. Bevacqua, E. de Sevin, C. Pelachaud, M. McRorie, and I. Sneddon, "Building credible agents: Behaviour influenced by personality and emotional traits," in *Proc. Kansei Engineering and Emotion Research*, Paris, France, 2010.

[18] D. Goddeau, E. Brill, J. R. Glass, C. Pao, M. Phillips, J. Polifroni, S. Seneff, and V. W. Zue, "Galaxy: A human-language interface to on-line travel information," in *Proc. ICSLP*, Yokohama, Japan, 1994.

[19] V. W. Zue and J. R. Glass, "Conversational interfaces: Advances and challenges," *Proceedings of the IEEE*, vol. 88, no. 8, p. 1166–1180, 2002.

[20] J. Allen, G. Ferguson, and A. Stent, "An architecture for more realistic conversational systems," in *Proc. Intelligent User Interfaces*, Santa Fe, USA, 2001, pp. 1–8. [Online]. Available: http://portal.acm.org/citation.cfm?id=359822

[21] O. Lemon and O. Pietquin, "Machine learning for spoken dialogue systems," in *Proc. of Interspeech*, 2007.

[22] A. Raux, "Flexible Turn-Taking for spoken dialog systems," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, USA, 2008.

[23] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in english and japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, Jul. 2000. [Online]. Available: http://dx.doi.org/10.1016/S0378-2166(99)00109-5

[24] M. Schröder, D. Heylen, and I. Poggi, "Perception of non-verbal emotional listener feedback," in *Proc. Speech Prosody 2006*, Dresden, Germany, 2006.

[25] D. Pardo, B. L. Mencia, A. H. Trapote, and L. Hernandez, "Non-verbal communication strategies to improve robustness in dialogue systems: a comparative study," *Journal on Multimodal User Interfaces*, vol. 3, no. 4, pp. 285–297, 2010. [Online]. Available: http://www.springerlink.com/content/k58267064t133350/

[26] W. Wahlster, "Smartkom: Symmetric multimodality in an adaptive and reusable dialogue shell," in *Proc. Human Computer Interaction Status Conference*, vol. 3, 2003, p. 47–62.

[27] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan, "Human conversation as a system framework: designing embodied conversational agents," in *Embodied conversational agents*. MIT Press, 2000, pp. 29–63. [Online]. Available: http://portal.acm.org/citation.cfm?id=371555

[28] S. Hyniewska, R. Niewiadomski, M. Mancini, and C. Pelachaud, "Expression of affects in embodied conversational agents," in *A blueprint for Affective Computing*, K. R. Scherer, T. Bänziger, and E. B. Roesch, Eds. Oxford University Press, 2010.

[29] M. Schröder, "Expressive speech synthesis: Past, present, and possible futures," in *Affective Information Processing*, J. Tao and T. Tan, Eds. London: Springer, 2009, pp. 111–126. [Online]. Available: http://dx.doi.org/10.1007/978-1-84800-306-4_7

[30] K. van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, and S. Baumann, "Fully generated scripted dialogue for embodied agents," *Artificial Intelligence*, vol. 172, no. 10, pp. 1219–1244, Jun. 2008.

[31] B. Kempe, N. Pfleger, and M. Löckelt, "Generating verbal and nonverbal utterances for virtual characters," in *Virtual Storytelling*. Springer, 2005, pp. 73–76. [Online]. Available: http://dx.doi.org/10.1007/11590361_8

[32] N. Krämer, L. Hoffmann, and S. Kopp, "Know your users! empirical results for tailoring an agent's nonverbal behavior to different user groups," in *Proc. Intelligent Virtual Agents*, Philadelphia, USA, 2010, p. 468–474.

[33] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual and spontaneous expressions," in *Proc. Multimodal Interfaces*, Nagoya, Japan, 2007, pp. 126–133. [Online]. Available: http://portal.acm.org/citation.cfm?id=1322192.1322216

[34] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous *et al.*, "Whodunnit - searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, 2011.

[35] F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber, "An emotion-aware voice portal," in *Proc. Electronic Speech Signal Processing ESSP*, Prague, Czech Republic, 2005, p. 123–131.

[36] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of Social Issues*, vol. 56, no. 1, pp. 81–103, 2000. [Online]. Available: http://dx.doi.org/10.1111/0022-4537.00153

[37] F. Biocca, J. Burgoon, C. Harms, and M. Stoner, "Criteria and scope conditions for a theory and measure of social presence," in *Presence 2001*, Philadelphia, 2001.

[38] A. von der Pütten, N. C. Krämer, and J. Gratch, "Who's there? can a virtual agent really elicit social presence?" in *Proc. 12th Annual International Workshop on Presence*, Los Angeles, CA, USA, 2009.

[39] J. Cassell, "Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents," in *Embodied conversational agents*. MIT Press, 2000, pp. 1–27. [Online]. Available: http://portal.acm.org/citation.cfm?id=371554

[40] V. Vinayagamoorthy, M. Gillies, A. Steed, E. Tanguy, X. Pan, C. Loscos, and M. Slater, "Building expression into virtual characters," in *Eurographics Conference State of the Art Report*, Vienna, Austria, 2006.

[41] E. André and C. Pelachaud, "Interacting with embodied conversational agents," in *Speech technology*, F. Chen and K. Jokinen, Eds. New York: Springer, 2010. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-73819-2_8

[42] D. Heylen, E. Bevacqua, C. Pelachaud, I. Poggi, J. Gratch, and M. Schröder, "Generating listening behaviour," in *Emotion-Oriented Systems – The Humaine Handbook*, P. Petta, C. Pelachaud, and R. Cowie, Eds. Springer, 2010, pp. 321–348.

[43] D. K. Heylen, "Head gestures, gaze and the principles of conversational structure," *International Journal of Humanoid Robotics*, vol. 3, no. 3, pp. 241–267, 2006.

[44] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, "Creating rapport with virtual agents," in *Proc. Intelligent Virtual Agents*, Paris, France, 2007, pp. 125–138. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-74997-4_12

[45] J. C. Acosta, "Using emotion to gain rapport in a spoken dialog system," PhD Thesis, University of Texas at El Paso, 2009.

[46] R. Porzel and M. Baudis, "The tao of CHI: towards effective human-computer interaction," in *Proc. HLT/NAACL*, Boston, MA, USA, 2004.

[47] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman, 1982.

[48] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[49] J. Weizenbaum, "ELIZA - a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966. [Online]. Available: doi:10.1145/365153.365168

[50] 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 *LREC2008 Workshop on Corpora for Research on Emotion and Affect*, Marrakech, Morocco, 2008, pp. 1–4.

[51] D. Heylen, A. Nijholt, and M. Poel, "Generating nonverbal signals for a sensitive artificial listener," in *Verbal and Nonverbal Communication Behaviours*. Springer, 2007, pp. 264–274. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-76442-7_23

[52] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme," in *SEMAINE Report D6b*, 2010. [Online]. Available: http://semaine-project.eu/D6b_labelled_data.pdf

[53] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not Two-Dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007. [Online]. Available: doi:10.1111/j.1467-9280.2007.02024.x

[54] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005. [Online]. Available: doi:10.1177/0539018405058216

[55] R. Reisenzein, "Pleasure-Arousal theory and the intensity of emotions," *Journal of Personality and Social Psychology*, vol. 67, no. 3, pp. 525–539, Sep. 1994.

[56] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': an instrument for recording perceived emotion in real time," in *Proc. ISCA Workshop on Speech and Emotion*, Northern Ireland, 2000, pp. 19–24.

[57] R. Cowie, C. Cox, J.-C. Martin, A. Batliner, D. Heylen, and K. Karpouzis, "Issues in Data Labelling," in *Emotion-Oriented Systems: The Humaine Handbook*, R. Cowie, C. Pelachaud, and P. Petta, Eds. Berlin, Heidelberg: Springer-Verlag, oct 2010, pp. 213–241.

[58] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[59] L. Devillers, R. Cowie, J. C. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie, "Real life emotions in french and english TV video clips: an integrated annotation protocol combining continuous and discrete approaches," in *Proc. LREC*, Genoa, Italy, 2006.

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

[61] 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," in *Proc. LREC workshop on Emotion Corpora*, Valettea, Malta, 2010, pp. 42–46.

[62] M. Johnston, P. Baggia, D. C. Burnett, J. Carter, D. A. Dahl, G. McCobb, and D. Raggett, "EMMA: Extensible MultiModal Annotation markup language," World Wide Web Consortium, W3C Recommendation, Feb. 2009. [Online]. Available: http://www.w3.org/TR/emma/

[63] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thórisson, and H. Vilhjálmsson, "Towards a common framework for multimodal generation: The behavior markup language," in *Proc. Intelligent Virtual Agents*, Marina del Rey, CA, USA, 2006, pp. 205–217. [Online]. Available: http://dx.doi.org/10.1007/11821830_17

[64] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in *Proc. of Interspeech*, Brighton, UK, 2009, pp. 1595–1598.

[65] F. Eyben, H. Gunes, M. Pantic, M. Schroeder, B. Schuller, M. F. Valstar, and M. Wollmer, "User-profiled human behaviour interpreter," in *SEMAINE Report D3c*, 2010. [Online]. Available: http://semaine.sourceforge.net/SEMAINE-3.0/D3c%20Human%20behaviour%20interpreter.pdf

[66] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "Recognition of spontaneous conversational speech using long short-term memory phoneme predictions," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 1946–1949.

[67] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[68] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework," *Cognitive Computation*, vol. 2, no. 3, pp. 180–190, 2010.

[69] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "COSINE - a corpus of multi-party conversational speech in noisy environments," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[70] P. Viola and M. Jones, "Robust real-time object detection," *Int J Comput Vis*, vol. 57, no. 2, pp. 137–154, 2002.

[71] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution grey-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[72] C.-C. Chang and C.-J. Lin, *LibSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[73] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1760–1774, 2009.

[74] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

[75] ——, "Automatic measurement of affect in dimensional and continuous spaces: Why, what, and how?" in *Proc. of Measuring Behavior*, 2010, pp. 122–126.

[76] M. Nicolaou, H. Gunes, and M. Pantic, "Audio-visual classification and fusion of spontaneous affective data in likelihood space," in *Proc. of IEEE Int. Conf. on Pattern Recognition*, 2010, pp. 3695–3699.

[77] M. ter Maat and D. Heylen, "Selecting appropriate agent responses based on non-content features," in *Proc. Affective Interaction in Natural Environments, ACM Multimedia*, Firenze, Italy, 2010.

[78] R. M. Maatman, J. Gratch, and S. Marsella, "Natural behavior of a listening agent," in *5th International Conference on Interactive Virtual Agents*, Kos, Greece, 2005.

[79] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in english and japanese," *Journal of Pragmatics*, vol. 23, pp. 1177–1207, 2000.

[80] I. Poggi, *Mind, hands, face and body. A goal and belief view of multimodal communication*. Berlin: Weidler, 2007.

[81] S. H. E. de Sevin and C. Pelachaud, "Influence of personality traits on backchannel selection," in *Proc. Intelligent Virtual Agents*, Philadelphia, USA, 2010.

[82] M. McRorie, I. Sneddon, E. de Sevin, E. Bevacqua, and C. Pelachaud, "A model of personality and emotional traits," in *Proc. Intelligent Virtual Agents*, Amsterdam, 2009.

[83] M. Mancini and C. Pelachaud, "Distinctiveness in multimodal behaviors," in *Conference on Autonomous Agents and MultiAgent System*, 2008.

[84] M. Schröder, S. Pammi, and O. Türk, "Multilingual MARY TTS participation in the blizzard challenge 2009," in *Blizzard Challenge 2009*, Edinburgh, UK, 2009.

[85] S. Pammi, M. Charfuelan, and M. Schröder, "Multilingual voice creation toolkit for the MARY TTS platform," in *Proc. LREC*. Valettea, Malta: ELRA, 2010. [Online]. Available: http://www.dfki.de/lt/publication_show.php?id=4878

[86] S. Pammi, M. Schröder, M. Charfuelan, O. Türk, and I. Steiner, "Synthesis of listener vocalisations with imposed intonation contours," in *Proc. Seventh ISCA Tutorial and Research Workshop on Speech Synthesis*. Kyoto, Japan: ISCA, 2010. [Online]. Available: http://www.dfki.de/lt/publication_show.php?id=4886

[87] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, "Greta: an interactive expressive eca system," in *AAMAS'09 - Autonomous Agents and MultiAgent Systems*, Budapest, Hungary, 2009.

[88] S. Westerman, P. Gardner, and E. Sutherland, "Usability testing Emotion-Oriented computing systems: Psychometric assessment," HUMAINE deliverable D9f, 2006. [Online]. Available: http://emotion-research.net/projects/humaine/deliverables/D9f%20Psychometrics%20-%20Final%20-%20with%20updated%20references.pdf

[89] R. Cowie, "Perceiving emotion: towards a realistic understanding of the task," *Philosophical Transactions B*, vol. 364, no. 1535, pp. 3515–3525, Dec. 2009. [Online]. Available: http://dx.doi.org/10.1098/rstb.2009.0139

[90] R. Cowie and E. Douglas-Cowie, "Prosodic and related features that signify emotional colouring in conversational speech," in *The Role of Prosody in Affective Speech Studies in Language and Communication*, S. Hancil, Ed. Berne: Peter Lang, 2009, vol. 97, pp. 213–240.

[91] P. M. Niedenthal, L. W. Barsalou, P. Winkielman, S. Krauth-Gruber, and F. Ric, "Embodiment in attitudes, social perception, and emotion," *Personality and Social Psychology Review*, vol. 9, no. 3, pp. 184–211, 2005. [Online]. Available: http://psr.sagepub.com/cgi/content/abstract/9/3/184

[92] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-Corpus classification of realistic emotions – some pilot experiments," in *Proc. LREC workshop on Emotion Corpora*, Valettea, Malta, 2010, pp. 77–82.

[93] M. Brendel, R. Zaccarelli, B. Schuller, and L. Devillers, "Towards measuring similarity between emotional corpora," in

Proc. LREC workshop on Emotion Corpora, Malta, 2010, pp. 58–64.
[94] J. Bachorowski, "Vocal expression and perception of emotion," *Current Directions in Psychological Science*, vol. 8, no. 2, pp. 53–57, 1999.
[95] ISO, "Ergonomic requirements for office work with visual display terminals (VDTs) - part 11: Guidance on usability," International Standards Organisation, ISO Standard ISO 9241-11, 1998.
[96] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: a framework for evaluating spoken dialogue agents," in *Proc. EACL*, Madrid, Spain, 1997, pp. 271–280. [Online]. Available: http://portal.acm.org/citation.cfm?id=979652&dl=
[97] M. Edwardson, "Measuring consumer emotions in service encounters: an exploratory analysis," *Australasian Journal of Market Research*, vol. 6, no. 2, p. 34–48, 1998.
[98] M. Csikszentmihalyi and I. S. Csikszentmihalyi, *Beyond boredom and anxiety*. San Francisco, USA: Jossey-Bass, 1975.
[99] M. V. Sanchez-Vives and M. Slater, "From presence to consciousness through virtual reality," *Nature Reviews Neuroscience*, vol. 6, no. 4, pp. 332–339, 2005. [Online]. Available: http://dx.doi.org/10.1038/nrn1651
[100] K. Isbister, K. Höök, M. Sharp, and J. Laaksolahti, "The sensual evaluation instrument: developing an affective evaluation tool," in *Proc. SIGCHI conf. on Human Factors in computing systems*, Montréal, Canada, 2006, pp. 1163–1172. [Online]. Available: http://portal.acm.org/citation.cfm?id=1124772.1124946
[101] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme," *SEMAINE project*, 2010. [Online]. Available: http://semaine-project.eu/D6b_labelled_data.pdf
[102] G. Breaden-Madden, "Emotionally coloured discourse: Non-verbal behaviours as indicators of communication breakdown," in *Proceedings of the British Psychological Society, Northern Ireland Branch Annual Conference*, Enniskillen, 2011.
[103] P. Baxter, T. Belpaeme, L. Cañamero, P. Cosi, Y. Demiri, and V. Enescu, "Long-term human-robot interaction with young users," in *Proc. Workshop Robots with Children at HRI 2011*, Lausanne, Switzerland, 2011.

Marc Schröder is a Senior Researcher at the DFKI LT lab. He has received his PhD in Speech Science from Saarland University in 2003. He is Coordinator of the FP7 project SEMAINE. Schröder acts as Portal Editor for the HUMAINE Association, as Editor for the W3C Emotion Markup Language, and as Associate Editor for the IEEE Transactions on Affective Computing.

Elisabetta Bevacqua is a postdoc at CNRS in Paris. She got her PhD in Computer Science at the University of Paris VIII in 2009. Her research field includes the verbal and non verbal communication, human-machine interaction and the implementation of models to simulate the humans' behaviour for virtual agents, particularly while listening to a user.

Roddy Cowie is Professor of Psychology at Queen's University Belfast. Recently he has focussed on emotion and computing through a series of projects funded by the European Union. He was first president of the HUMAINE Association for emotion-oriented computing and chairs the Steering Board of the IEEE Transactions on Affective Computing.

Florian Eyben is a third year Ph.D. student at the Institute for Human-Machine Communication at the Technische Universität München, where he is currently working on audio feature extraction and affect recognition from spontaneous conversational speech within the FP7 project SEMAINE.

Hatice Gunes received her Ph.D. degree in Computer Science from University of Technology Sydney (UTS), Australia, in 2007. She is currently a Lecturer at Queen Mary University of London. She is a (co-)organizer of the EmoSPACE Workshop at IEEE FG 2011 and a member of the Editorial Advisory Board for the Affective Computing and Interaction Book (IGI Global, 2011).

Dirk Heylen studied linguistics, computer science and computational linguistics at Antwerp University. His current research at the University of Twente concerns conversational informatics which involves analysing human-human conversations automatically and creating agents and robots that have some conversational skills.

Mark ter Maat is a PhD student at the Human Media Interaction group at the University of Twente. For his Masters Thesis in the same group, he worked on a Virtual Conductor. Right now he is working on dialogue systems, and more particularly on turn taking and behaviour selection.

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

Sathish Pammi is a Researcher and PhD student at DFKI. Since 2008, he has been working on the development of Text-To-Speech (TTS) systems, including synthesis of vocal listener behavior, for Sensitive Artificial Listeners (SAL).

Maja Pantic is Professor in Affective and Behavioural Computing at Imperial College London (Computing), UK, and at the University of Twente (EEMCS), NL. She is one of the world's leading experts in the research on machine analysis of human behavior including facial expressions, body gestures, human affect and social signals.

Catherine Pelachaud is Director of Research at CNRS in the laboratory LTCI, TELECOM ParisTech. She received her PhD in Computer Graphics at the University of Pennsylvania in 1991. Her research interest includes embodied conversational agents, representation languages for agents, nonverbal communication, expressive behaviors and multimodal interfaces.

Björn Schuller received his diploma in 1999 and his doctoral degree in 2006, both in electrical engineering and information technology from TUM in Munich/Germany where he is presently tenured as Senior Researcher and Lecturer in Pattern Recognition and Speech Processing. Dr. Schuller has (co-)authored more than 180 peer reviewed publications.

Etienne de Sevin after a master degree in cognitive sciences, received a PhD degree in Computer Science from VRLab EPFL, Switzerland in 2006. His research interest focus on action selection in real-time of autonomous virtual agents.

Michel Valstar (PhD 2008) is a Research Associate in the intelligent Behaviour Understanding Group (iBUG) at Imperial College London, Department of Computing. Currently he is working in the fields of computer vision and pattern recognition, specialising in the analysis of facial expressions.

Martin Wöllmer works as a researcher at the Technische Universität München (TUM). He obtained his diploma in Electrical Engineering and Information Technology from TUM. His current research focus lies on multimodal data fusion, automatic recognition of emotionally colored and noisy speech, and speech feature enhancement.