# 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 bring some 'soft skills' to a real-time interactive multimodal dialogue system. Our aim is to provide technology which exhibits a certain amount of competence in interpreting and producing non-verbal behaviour that helps sustain a conversational dialogue. We focus on the Sensitive Artificial Listener (SAL) scenario, a type of dialogue that mainly relies on the analysis and generation of non-verbal behaviour, and that requires only extremely limited verbal understanding on the part of the machine. We motivate how this scenario allows us to concentrate on non-verbal capabilities without running into insurmountable obstacles of spoken language understanding. We describe the integrated real-time system we created, which combines incremental analysis of user behaviour, dialogue management, and synthesis of speaker and listener behaviour of an Embodied Conversational Agent (ECA). 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, allowing for systematic and comparative investigation of a multitude of research questions. We present initial evaluation of individual system components, and discuss in some depth principles that according to us should underlie the evaluation of SAL-type systems.

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

## 1 Introduction

Naturally talking with a machine in the same way as we talk with one another is a distant goal which is not achievable in the short term. Many things remain to be done to achieve this goal; for example, much effort has been, and continues to be invested in high-quality Automatic Speech Recognition (ASR) [1]; in designing dialogue structures and modelling task domains such that a dialogue goal can be achieved efficiently [2]; or in the grounding of a machine's knowledge about the world, so that human and machine have a common experiential basis for the interaction [3].

Complementary to these topics, the present paper addresses the issue of sustaining the conversation itself, the 'soft skills' required to smoothly talk to one another. What does it take to 'manage' a conversation so that it feels natural? What social and emotional aspects are needed to maintain enough of a relation to continue chatting?

Numerous elements contribute to the successful management of a conversation. Achieving a sense of 'social presence' depends on co-presence, psychological involvement, and behavioural engagement [4]. A spoken interaction requires a range of capabilities. Obviously, some knowledge of turn-taking is needed [5] – of knowing when to speak and when to be silent. Even when listening, though, it is necessary to act appropriately – to signal that one is still present and participating in the conversation, and to provide feedback as to how the speaker's message is being received [6]. Throughout the speaker and listener roles, a participant in a conversation is expected to display a consistent image – of one's own state, of the relation with the interlocutor [7], and of the topic of conversation. A machine that would implement all these capabilities perfectly should be perceived as a fully natural conversational partner by a human user. Given that this is not possible at this stage, the interesting question is which approximations are sufficient to achieve *some* sense of naturalness.

Relevant questions include for example the following. To what extent is verbal understanding necessary to achieve a sense of rapport? Which aspects of non-verbal user behaviour are important to pick up? Which aspects of system behaviour is it important to get right – timing, type of expression, intensity of an expression, etc.? To what extent do these depend on the context of what was previously said and done? Are there 'default' behaviours that can be used when

- *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, H. Gunes and M. Valstar are with Imperial College of Science, Technology and Medicine, 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.*

no reliable information about the context is available? What happens when a wrong perception triggers inadequate system behaviour? Are there promising repair strategies for such cases?

These and many more questions should be answered in order to develop different elements of naturalness in human-machine conversation. Much can be learned from the observation of human-to-human conversations. However, we believe that one of the best ways to understand a system is trying to build it. Therefore, in the present paper we report on our attempt to build a fully autonomous multimodal dialogue system with some non-verbal capabilities. This work has a double aim. On the one hand, we explore what level of naturalness we can achieve at this stage of technological development. On the other hand, the resulting system can be used as a versatile test environment for research on natural human-machine interaction.

### 1.1 Related work

A number of full-scale interactive systems have been created which take emotional aspects into account.

Several endeavours have attempted to embed emotional factors into games and entertainment scenarios. For example, the NECA project [8] generated scripted dialogues between embodied conversational agents 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 [9] supported dialogues involving multiple humans and multiple agents. Both humans and agents were represented in the system by Conversational Dialogue Engines [10] communicating with each other using the concepts of an application-specific ontology. Emotions to be expressed by the characters were computed by rules operating on the domain knowledge, and were realised through facial expression, skin texture, tears, and breathing patterns. The project IDEAS4Games realised an emotion-aware poker game, in which two agents and a user played against each others with physical cards carrying RFID tags [11]. The emotions of characters were computed from game events using an affective reasoner [12], and realised through the synthetic voice and through body movements. Castellano et al. [13] built an iCat robot that plays chess with children and expresses emotions through facial expressions after the child has made a move. In all these scenarios, the emotion is determined by the game or application logic, and expressed through the synthetic characters. User emotions are not taken into account.

In the area of e-learning, the role of student emotions is recognised [14], but few systems appear to exist which model emotions explicitly. One such system is *FearNot!* [15], an educational application helping children to deal with bullying. The system uses an architecture involving reactive and deliberative layers and memory components, as well as sensors and effectors. The student is presented with a virtual story involving several pupils involved in bullying situations, and needs to help the victim learning to deal with the situation by exploring various possible actions. The emotions consistent with the virtual characters' roles are displayed through cartoon-style visual rendering and the characters' voices.

Automatic analysis of human emotions has been applied, e.g., in a voice portal [16]. The idea is to detect a customer's anger in time to redirect the customer to a human agent. A key problem in this task is the trade-off between false negatives (not detecting an angry customer) and false positives (treating a customer as angry who is not angry). Another application is surveillance, as in the automatic detection of fear [17] in call centres or public spaces.

Works addressing conversation directly as the main task are relatively few. The Rapport agent [7] observes the head movements and voice prosody of a user telling a story, and generates contingent visual listener behaviour including nods and posture shifts. In a carefully controlled study, the automatically generated listener behaviour was rated similarly well as natural human listener behaviour in a face-to-face condition, and significantly better than a non-contingent version of the system (effectively playing back the contingent behaviour generated from the *previous* subject). The rapport agent produces no vocal feedback, and it never speaks. The effectiveness of contingent emotional adaptation to a user's emotion in a dialogue system was investigated by Acosta [18]. In a speech-based dialogue system aiming to persuade students of the values of graduate school, the emotion-related prosody of the system's utterances was modified as a function of the emotion recognised from the preceding user utterance. System utterances were generic phrases that followed a pre-defined script; only their prosody was adapted. Users rated the system as significantly better on a number of rapport-related scales, compared to a neutral baseline as well as a non-contingent version where the expressivity matched the previous rather than the current subject.

To the best of our knowledge, no full-scale dialogue system has been built before that takes into account the user's emotion from visual and vocal non-verbal cues, and interacts in real time both as a speaker and a listener in a multimodal conversational setting. Furthermore, none of the integrated systems mentioned above is publicly available as open source, which makes it difficult to improve existing work incrementally.

A real-time interactive emotion-oriented system depends on solutions to difficult problems in many areas (cf. [19]). Due to lack of space, we merely refer to articles reviewing the state of the art in the following areas: analysis of user behaviour from face [20] and voice [21]; understanding the role of non-verbal behaviour in conversation [22]; taking into account non-verbal user behaviour in dialogue planning [23]; non-verbal behaviour of speakers [24], [25] and listeners [26]; generation of expressive synthetic speech [27] and facial and bodily behaviour [28].

## 1.2 The Sensitive Artificial Listener scenario

The Sensitive Artificial Listener (SAL) is a scenario for human-machine dialogue that is designed to allow for the study of non-verbal aspects of conversation – the *soft skills* – without requiring extensive verbal understanding or task intelligence on the part of the machine. It draws on the idea of Weizenbaum's original ELIZA system [29], a text-based chat system that engaged users by encouraging them to talk more, using stock phrases and follow-up questions. ELIZA used simple textual pattern matching to analyse the user's input.

While SAL is like ELIZA in some ways, it is diametrically opposite in others. 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. 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 gaze, head movements, facial expression, and voice. However, the competences that let it use these are not to do with language at all. Spoken words are simply treated as sounds that express emotion.

The rationale behind SAL is explained in Section 2. However, it is useful to introduce what a user would encounter. The description here compresses a series of versions of increasing sophistication [30], [31], [32] (also described in Section 2). The user encounters four 'Sensitive Artificial Listeners', shown in Fig. 1. Each has a different personality, and each tries to 'pull' the user towards its own emotional state. The four characters reflect the four quadrants in arousal-valence space. Spike is aggressive, confrontational, and enjoys an argument; Poppy is cheerful, optimistic, and looks on the bright side of life; Obadiah is gloomy, and has a pessimistic outlook; Prudence is matter-of-fact, and has a practical view on life.

Each SAL's utterances are restricted to a script of predefined phrases, which are used to introduce topics and to encourage the user to follow-up on a topic. The phrase that it selects 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).

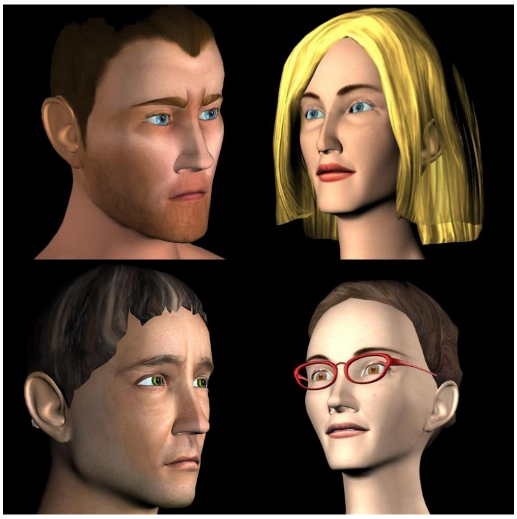


Fig. 1. The four SAL characters as they appear in automatic SAL: aggressive Spike; cheerful Poppy; gloomy Obadiah; and pragmatic Prudence.

The SAL scenario has repeatedly shown to work well with users who are willing to engage with the system; 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. Insofar, an 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. As such, it is a fruitful environment for investigating the non-verbal capabilities required to sustain a human-machine conversation.

## 1.3 Scope of the present paper

The present paper provides a technically oriented view on the autonomous Sensitive Artificial Listener system. Having motivated why we consider this to be a promising framework, we first explain the reasoning that has led to the SAL scenario (Section 2), and describe several human-driven versions of SAL which served to understand relevant variables and to collect data for use in quantitative and qualitative analyses. We then provide an encompassing account of the autonomous SAL system (Section 3), describing its design principles, architecture, and the capabilities and limitations of the individual system components. We discuss principles for evaluating systems such as ours (Section 4) but reserve an extensive evaluation study for a future publication. The paper concludes with a discussion of the specific contribution made by the present work to the progression of the Affective Computing research area.

## 2 Sensitive Human Listeners

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 [33] 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.

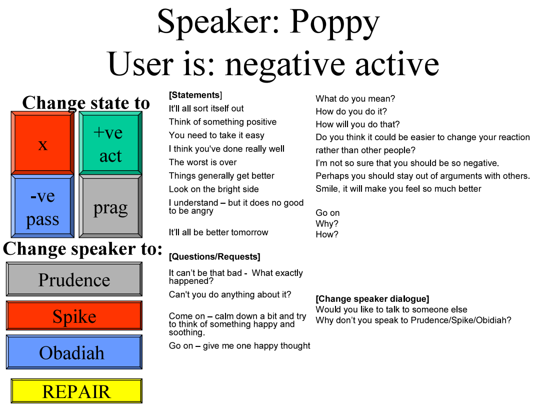
The SAL scenario was invented explicitly and deliberately with a view to setting a useful level of challenge. The idea was prompted by work with chat shows, where hosts appeared to follow a strategy that was simple and effective: register the guest's emotions, and throw back a phrase that gives very 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 'soft skills' without being distracted by difficult problems in speech recognition and natural language processing.

The studies reported in this section have a dual function. On one side, they are concerned with checking that a situation of that general kind can be created, and does actually have the interesting properties that it seems to. On the other, they are concerned with acquiring the data that is needed to model the relevant soft skills. The two are logically separate, but practically intertwined: recordings of tests provide the data.

The first key step was to establish that the supposed human ability did in fact exist – that is, that human beings could truly sustain an emotionally coloured interaction where one party (the 'host') used only stock phrases chosen to provoke emotional reactions. In early attempts, conversation ran dry very quickly. The structure which was devised to counter that has stayed constant since. To introduce variety, the single 'host' was replaced by four contrasting 'artificial listeners'. To create a sense of cohesion, each artificial listener had a consistent personal style, and a distinct agenda. These had to be definable in terms of the abilities that were being considered, that is, detection and expression of emotion. Hence, each artificial listener had a default emotional state, and its agenda was to draw the other party (the 'user') into the same state. As a result, the 'artificial listeners' described earlier were defined, corresponding to regions of emotional space that it seemed likely existing technology could recognise (and hence decide whether the user was in the listener's favoured state). Spike was angry, and tried to make the user equally angry; Poppy was effervescently happy, and tried to make the user equally happy; Obadiah was despondent, and tried to make the user equally gloomy; Prudence was matter-of-fact, and tried to make the user equally matter-of fact. The characters tend to catch people's imagination, but it is worth stressing that the paradigm is not defined by the four listeners. They cater to the limitations of the emotion-oriented technologies available. It is a natural goal to extend the range of characters as the technologies improve.

The first stable version of SAL consisted of a dynamic script in which the options available at any given time depended on the 'listener' who was in play and the user's emotional state. An operator simulated the 'listener' by reading, with appropriate emotional colouring, one of the options offered by the script. Fig. 2 illustrates the appearance of a script at one particular moment. It was implemented in Powerpoint, which allowed the operator to move around the script by pressing the buttons on the left hand side – the top group when the user's state changed, the bottom group when the user asked to speak to a different listener. For that reason, this version has been dubbed



Speaker: Poppy
User is: negative active

[Statements]
It'll all sort itself out
Think of something positive
You need to take it easy
I think you've done really well
The worst is over
Things generally get better
Look on the bright side
I understand – but it does no good to be angry
It'll all be better tomorrow

What do you mean?
How do you do it?
How will you do that?
Do you think it could be easier to change your reaction rather than other people?
I'm not so sure that you should be so negative.
Perhaps you should stay out of arguments with others.
Smile, it will make you feel so much better

Go on
Why?
How?

[Questions/Requests]
It can't be that bad - What exactly happened?
Can't you do anything about it?
Come on – calm down a bit and try to think of something happy and soothing.
Go on – give me one happy thought

[Change speaker dialogue]
Would you like to talk to someone else
Why don't you speak to Prudence/Spike/Obidiah?

Fig. 2. Example phrases used in the SAL framework.

'Powerpoint SAL'.

The work with Powerpoint SAL confirmed that people can indeed engage in interactions of this kind for sustained periods. 24 sessions were recorded, with 20 different users: the average session lasted over 20 minutes. The emotional content of the material was described using the FEELTRACE system [34], which yields descriptions in terms of the two most widely used affective dimensions, valence (how positive or negative the person feels) and arousal (how active or lethargic the person feels). A few users struggled to engage, but most responded with what raters judged was genuine emotion, often quite intense.

The implication is that if a machine could match the skills that the operator brought to the interactions, it would be able to sustain similar conversations. The skills in question appear to be both significant for human-machine interaction, and quite restricted. They include recognising the user's emotional state; giving appropriate expression to the scripted utterances; basic conversational skills (involving turn-taking and backchanneling); and selecting the appropriate item from the kind of menu shown in Fig. 2. Critically, these need to be co-ordinated, and to be executed in real time. Capturing that constellation of skills is a task that seems a worthwhile challenge.

The studies using Powerpoint SAL generated a substantial quantity of data. The audio data has been used in a number of projects on emotion recognition [35], [36]. Video was recorded, but work on it remains unpublished.

The next step in development was a 'wizard of Oz' system known as 'Semiautomatic SAL'. The main difference between it and Powerpoint SAL was that instead of the operator reading the relevant utterances, he/she clicked on them, and a recorded version of the utterance, in a voice suited to the character, was played. Semiautomatic SAL is a natural step towards automation, facilitates high quality recording, and makes it easier to apply experimental manipulations.

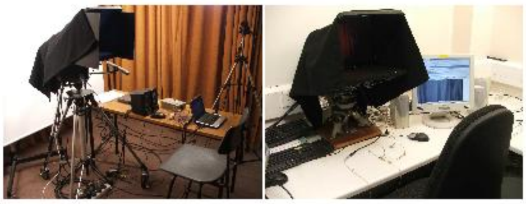


Fig. 3. Recording setup for user (left) and operator (right) as used with Semiautomatic SAL and Solid SAL.

One of the key aims of the work with Semiautomatic SAL and later versions was to capture data that would support automatic analysis in both audio and video modalities. To do that, a specialised recording configuration was developed. It is shown in Fig. 3. 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. That makes it possible for each party to have the impression of looking directly into the face of the other, and at the same time to be filmed by cameras pointing directly at his/her face. The setup also included multiple microphones for each participant. A specially designed computer with multiple hard disc drives was needed to capture data from all these sources.

The first priority with this system was to verify that the task being set for the automatic system was achievable. The automatic system would be required to choose responses on the basis of almost wholly nonverbal cues – facial expression, vocal signals, and so on. In contrast, the Powerpoint SAL operator had been able to use the user's words to guide his/her choice of response. Because the Semiautomatic SAL system separated user and operator, it became straightforward to check whether the verbal component was essential. Pilot experiments used an acoustic filter technique, but they suggested that the point could be made even more strongly using a simple manipulation, which was to compare a condition where the operator could both see and hear the user with a condition where the auditory channel from user to operator was removed completely. Interactions were then recorded with twelve users. Each user interacted with all four characters: the operator could hear the user in two cases, and not in the other two (combinations of characters with and without sound were balanced across users). The interactions included a battery of tests developed to assess the quality of the interaction. They are described in Section 4. The key point at this stage is that user ratings of interaction quality did not decline when the operator had no sound, confirming that people have resources that allow them to choose appropriate responses without verbal information.

The work with Semiautomatic SAL provided high quality recordings with some distinctive features, not least breakdowns of communication. These are likely to be a feature of human-computer interaction for the foreseeable future, and it is important to have data that provides a basic for recognising them. Their form is discussed in Section 4. However, there are important kinds of behaviour that cannot be studied with Semiautomatic SAL, most obviously backchanneling (in a broad sense of the term) by the 'listener'. That is clearly important for an agent's ability to keep a human engaged in an interaction, but it cannot be displayed by a 'wizard' operating a Semiautomatic SAL system (he/she is likely to spend 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 understand the characters of the 'listeners', and tried to speak as the relevant character would do. A substantial body of recordings was collected in the Solid SAL configuration, totalling 8 hours and 37 minutes. It is available via the SEMAINE website (http://www.semaine-db.eu). On an informal level, the material has guided work on backchanneling in automatic versions of SAL. On a formal level, Cowie et al. [37] have reported statistical analyses of the head movements involved in backchanneling during Solid SAL interactions. Complex accounts of these movements have been put forward [38]. In practice, it seems that all but a few convey primarily emotional messages, which can be summed up in terms of the party's energy levels (how active/passive he/she feels) and valence (how positive/negative he/she feels). Details are in [37].

Labelling of the recordings from Semiautomatic SAL and Solid SAL used an extension of the trace paradigm used for Powerpoint SAL. All of the material was annotated with the two dimensions used previously, and three others that psychological evidence suggests are key to capturing affective colouring in general: Intensity (which is the longest established dimensions of all), along with Power and Anticipation/Expectation, which emerged from an influential recent study [39]. After rating on those five mandatory dimensions, raters identified optional descriptors which appeared to be particularly relevant to each interaction, and rated from moment to moment how well it described what they were seeing and hearing. Table 1 shows the optional descriptors, and the frequency with which each was used in the first phase of labelling solid SAL interactions. It can be seen that 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 (amusement, expresses agreement/disagreement, gives information, gives opinion, thoughtful, (not) at ease, happy).

The labellings provide a rich description of both the emotional colouring of interactions and what Baron-Cohen [40] has called affective-epistemic states, that is, states which involve both knowledge and feeling. They open the way to several lines of research. One is to measure interdependences between them, and therefore the extent to which they can be reduced. A second is to establish how well they can be inferred from features (in visual and audio modalities) that can be measured automatically. A third, which is where the SAL paradigm brings a unique dimension, is to evaluate how performance in the relevant areas affects interaction.

The target for Solid SAL labelling is that all 25 user sessions will be rated by six labellers, and the labellings will be made available with the recordings via the SEMAINE website. At the time of writing, over 100 of the 150 rating sessions have been completed, and most are on the website.

It is not accidental that the studies reported in this section are not complete. It reflects one of the main features of the research in the SAL paradigm. It invites an iterative process, in which computational research identifies interesting problems, psychologists identify portions of human behaviour with the potential to illuminate them, computational research provides the tools to explore them, psychological research analyses them in more depth, and so on. The ideal is obvious enough, but there are relatively few instances where it has been translated into practice.

# 3 Sensitive Artificial Listeners

The implementation of the SAL paradigm as an autonomous real-time multimodal system is a challenge on multiple levels. We report on our approach to integrating the system, conceptually and technically, and describe our solution to the implementation of the different system components.

## 3.1 Building an integrated system

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

### *3.1.1 Conceptual framework*

The conceptual architecture that orients the implementation of the SAL system is shown in Fig. 4. While the details are grossly simplified, the figure shows the main items. First, user behaviour is observed through a camera and a microphone, and low-level features are computed using a battery of *feature extractor* components. Features are low-level data computed from the raw signals, and are typically computed at a constant frame rate, e.g. every 10 ms for audio data, and every video frame for video data. These features are used by *analyser* components to compute an estimate of the

| Basic Emotions | Epistemic States | Interaction Process Analysis | Validity |
|---|---|---|---|
| 10 anger | 23 (not) certain | 9 shows solidarity | 11 breakdown of engagement |
| 2 disgust | 79 (dis) agreement | 15 shows antagonism | 0 anomalous simulation |
| 82 amusement | 22 (un) interested | 12 shows tension | 19 marked sociable concealment |
| 27 happiness | 39 (not) at ease | 14 releases tension | 5 marked sociable simulation |
| 21 sadness | 41 (not) thoughtful | 6 makes suggestion | |
| 10 contempt | 9 (not) concentrating | 2 asks for suggestion | |
| | | 42 gives opinion | |
| | | 3 asks for opinion | |
| | | 72 gives information | |
| | | 3 asks for information | |

TABLE 1
Optional descriptors, with the number of times each was used in the rating of Solid SAL interactions.

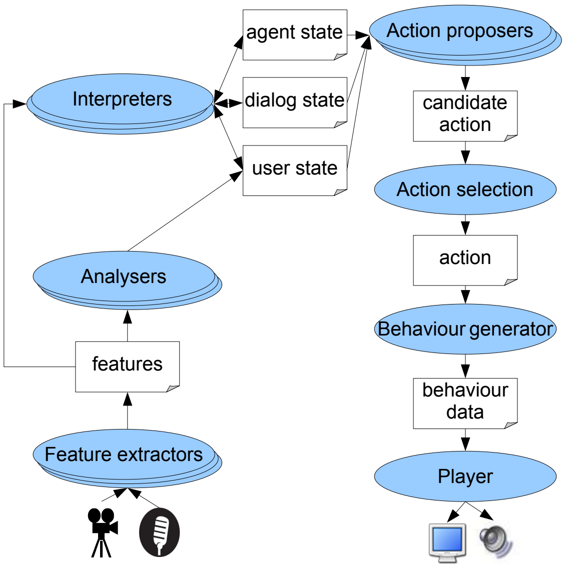


Fig. 4. Conceptual architecture of the SAL system (simplified).

current user state. We call *analysers* such components that try to make some sense of user behaviour without using context information, such as classifiers. The raw features and the preliminary estimate of the user state are further interpreted in the light of all available information by a set of *interpreter* components. These take decisions about the system's 'current best guess' about the state of the user, the dialogue and the agent. Interpreters do such diverse things as conclude when the user is speaking or not, make a final estimate of the current user emotion, and update the agent state such as the agent's degree of urgency to speak.

In parallel to this analysis and interpretation of the user's input, a group of *action proposers* continuously take decisions on whether to propose an action given the current state information. These include the action to speak, i.e. to produce a verbal utterance, as well as the action to exhibit some listener behaviour, such as a feedback signal or a mimicry backchannel. An *action selection* component makes sure only one action is being realised at a time. The selected action is then rendered in terms of concrete vocal, facial and gestural behaviour, and finally rendered by a *player* component.

All components are described in some detail below.

### 3.1.2 Technical framework for component integration

We have created a custom, cross-platform component integration framework, the SEMAINE API [41]. 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. We have chosen the message-oriented middleware ActiveMQ [42] as the remote communication layer, since it is reasonably fast and supports multiple message data formats including binary data. The SEMAINE API provides an abstraction, for Java and C++, in terms of `Components` that can `react()` to incoming messages and `act()` based on an internal timer.

The communication architecture is shown in Fig. 5. 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, representing the interconnection between components, their status and recent activity, a configurable selection of log messages, and optionally the messages being sent. The latter two functionalities are realised through a centralised logging mechanism, into which on the one hand every component can write, and on the other hand a *message logger* sends copies of messages being sent between components.

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 uses standard representation formats where possible. For example, the Extensible Multimodal Annotation (EMMA) language [43] from the World Wide Web Consortium (W3C) is used for representing the output of *analysers*; the Behaviour Markup Language

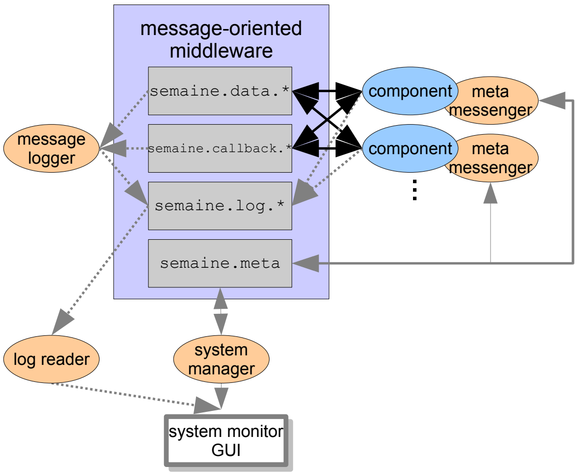


Fig. 5. Architecture of the SEMAINE API component integration framework

| Descriptors | Functionals |
|---|---|
| Intensity | position of max./min. |
| Zero-Crossing Rate | arithmetic mean |
| MFCC [1-12] | std. dev., skewness, kurtosis |
| Prob. of Voicing | centroid, duration |
| $F_0$ | quartiles & IQRs |
| $F_0$ Envelope | Percentile 5/95/98 |
| MFB 0–25 | Percentile-range 5–95 |
| LSP 0–7 | lin. regression coeff. 1/2 |
| | lin. regression error Q/A |
| | up-level time 75 & 90 |
| | down-level time 90 |
| | rise/fall time |
| | left/right curve time |

TABLE 2
Set of 3 060 acoustic features for affect recognition: 51 low-level descriptors with delta regression coefficients, 30 functionals. Abbreviations: MFCC: Mel-Frequency Cepstral Coefficients, ZCR: Zero-Crossing Rate, MFB: Mel-Frequency band, LSP: line spectral pairs, IQR: Inter-Quartile Range, Q/A: quadratic/absolute.

(BML) [44] is used for representing the SAL agent's behaviour in the process of realising actions, etc.

The *states* of user, agent and dialogue, on the other hand, are too domain-specific to be represented in a standard markup language. Instead, we have developed a flexible way of introducing domain-specific pieces of state information into the system. For each state information item, a config file links a unique short name (e.g., `headGesture`) to a namespace-aware XPath expression [45] (e.g., `/semaine:user-state/bml:bml/bml:head/@type`) which defines how the information is to be encoded in XML. The SEMAINE API supports state messages such that components can access the information via its short name both in reading and writing; the API knows from the config file how to encode and decode the information for transmitting it as a message. In our message-oriented middleware framework, there is no such thing as a jointly accessible memory; therefore, any state update by one component must be transmitted to all potential consumers of that information. We decided to keep track of the up-to-date state information locally within the API layer of each component rather than in a single *memory* component. In the trade-off between redundancy and efficiency, this has the advantage that, at the moment when the state information is requested, it is immediately available locally. Getting the information from a dedicated memory component would have required a question-answer pair of messages every time a piece of information is requested.

More details on the SEMAINE API, and examples for building new emotion-oriented systems in this framework, can be found in [41].

## 3.2 Feature extraction

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

### 3.2.1 *Acoustic features*

We describe the acoustic features used for analysis of user affect, prosody, non-linguistic vocalisations, voice activity detection, and keyword spotting in the following three subsections.

**Features for affect recognition.** The affect recognition module (see Section 3.3) is based on the TUM openSMILE feature extractor from the Munich open-source Emotion and Affect recognition toolkit (openEAR) [46]. A large set of 3 060 acoustic features is extracted from the audio signal. The feature space is constructed by applying 30 functionals to 51 acoustic low-level descriptor (LLD) contours of a certain length and their 51 first order delta regression coefficient contours. The functionals and low-level descriptors are listed in Table 2. A more thorough description of the individual features can be found in the documentation of the openSMILE feature extractor [47]. The length of the feature contours used for computing a single functionals vector is dynamically adaptive to the length of the user's speech turn. These feature vectors are computed incrementally every 0.5 seconds during segments of user speech (using all data from the beginning of the turn up to the current time); a final feature vector encompassing the whole turn is sent at the turn end. This enables incremental analysis of user affect while the user speaks, followed by a final refinement at the end of the turn. We describe the analysis modules (classifiers) for various dimensions of affect in Section 3.3.

**Features for prosodic analysis.** Our system uses basic prosodic features for detecting the end of a user speech turn and for determining the appropriate position for backchannel utterances. These features are contained in the set of acoustic LLD for affect recognition (Table 2), thus their extraction does not require extra computing resources. The descriptors which are used in particular are the pitch features $F_0$, $F_0$ envelope, and probability of voicing, and the signal intensity. These features are sent periodically every 10 ms.

**Voice activity.** In order to extract meaningful acoustic features only in regions where the user is talking, a voice activity detection (VAD) is used. The VAD uses a nearest neighbour classifier and Mel-frequency cepstral coefficients (MFCC) 0–12 to build a model of the speaker's voice. This simple classifier allows for very fast and efficient on-line model updates. If the VAD classifier output changes from 0 (silence/noise) to 1 (user speech) and remains at 1 for at least two frames, a 'speaking started' message is sent. When the voicing decision drops from 1 to 0 and remains at 0 for at least two frames a 'speaking stopped' message is sent.

**Keywords and non-linguistic vocalisations.** The system has to face a large number of out-of-vocabulary words. Therefore, the dialogue management component only uses 140 selected keywords for its decisions, which are detected by a tri-phone based keyword spotter. This enables limited verbal input to the system. Acoustic models were trained on a number of different speech corpora: the SAL corpus [48], the SEMAINE database [32], the Wall Street Journal (WSJ) corpus, as well as the AMIDA [49] and the AVIC database [50]. Non-linguistic vocalisations are included in these corpora (except WSJ), which are used to train 'phoneme' models for these vocalisations as investigated in [51]. The current best keyword hypothesis is sent continuously as an EMMA message, including keyword start times and confidences.

### 3.2.2 *Visual features*

The visual features extracted from the video signal include face detection and location, 2D-head motion, and facial point location. All visual features are sent with a frequency equal to the rate of frame capture.

**Face Detection.** The face detector used is based on the OpenCV implementation of the Viola & Jones face detector [52]. To ensure that the face detector component takes as little time as possible, we constrain the search space based on the maximum possible velocity of the head in a conversational scenario. The maximum expected head velocity towards and away from the camera places an upper and a lower bound on the change in size of the face in the next frame. Similarly, the maximum lateral and vertical velocity of the head with respect to the camera determine the maximum search area.

**2D-Head Motion Extraction.** In order to determine the magnitude and the direction of the 2D head motion, the optical flow is computed between two consecutive frames. It is applied to a refined face region (i.e., resized and smoothed) 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 $v = [v_x, v_y]$ represents the global horizontal and vertical head movement, as it takes into account all pixels within that region. After preliminary analysis, the angle feature $atan(v_y/v_x)$ has been considered as the most distinguishing feature to represent nods and shakes. The angle measure has then been discretised by representing it with directional codewords. The directional codeword is obtained by quantising the direction into four codes for head movements (for rightward, upward, leftward and downward motion, respectively) and one for 'no movement'.

**Facial Point Localisation.** We have developed and implemented a novel facial point detector that detects 20 fiducial facial points and the pupils in a near-frontal face. The method, coined BoRMaN [53], employs Boosted Regression and Markov Networks to predict the location of a target point $t$ relative to a randomly sampled patch centre location $L$. Thus each point in the neighbourhood of the target location can provide a prediction of where the target point is. This means we only need to try a small number of locations in order to get an accurate prediction $t$. The regressors we used were Support Vector Regressors. They were trained using the most informative Haar-like features, as selected by an implementation of Drucker's AdaBoost regressor [54].

The search space is constrained using Markov Random Field Networks (MRFs) that model the spatial relations between groups of points. These groups are defined as sets of points that have strong spatial constraints on each other even in the presence of facial expressions. For each group of points we have trained a separate MRF, and a separate MRF models the spatial relations between the group centres. In our implementation of the networks, the nodes do not represent the facial point locations themselves but instead each node models the relations between two points. The relations between nodes being modelled are their relative orientations and distances.

The point detector has been specifically trained to be able to cope with subjects of varying ages, sex, and ethnicity, and is able to cope well with various facial expressions, glasses, and facial hair. We evaluated the performance of the detector in terms of detection accuracy on images from three databases: the MMI Facial Expression Database [55], the FERET database [56], and the BioID database [57].

Not counting the chin point, in a 10-fold cross validation test on 400 images taken from the MMI Facial Expression and FERET databases 91.8% of the points were detected correctly. The chin point is detected less accurately, with an average error of 20% of the interocular distance, i.e. the distance between the pupils. In a second test the point detector was trained on the 400 images of the MMI and FERET database and tested

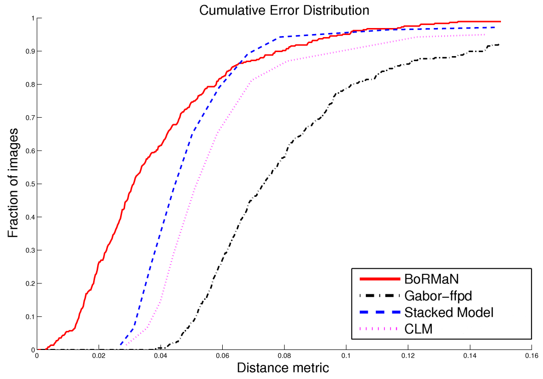


Fig. 6. Comparison of the cumulative error distribution of point to point error measured on the BioID test set.

| MSE | aro | exp | int | pow | val |
|---|---|---|---|---|---|
| Voice, SD | 0.069 | 0.064 | 0.089 | 0.107 | 0.108 |
| Head, SD | 0.076 | 0.064 | 0.082 | 0.101 | 0.104 |
| Voice, SI | 0.086 | 0.076 | 0.077 | 0.132 | 0.116 |
| Head, SI | 0.087 | 0.076 | 0.097 | 0.131 | 0.118 |

TABLE 3
Mean square error (MSE) of 5D affect prediction (arousal (aro), expectation (exp), intensity (int), power (pow), and valence (val)) experiments from head gestures (Head) and acoustic features (Voice) for subject dependent (SD) and subject independent (SI) evaluation splits.

on the BioID database. This allows a fair comparison with three state of the art point detectors (see Fig. 6), which have presented results on the same data set. The results showed that BoRMaN significantly outperforms the current state of the art.

### 3.3 Understanding human behaviour

This section describes the algorithms and methods used to analyse human behaviour based on acoustic and visual features as described in Section 3.2.

#### *3.3.1 Affect analysis from acoustic features*

The SAL system contains speech-based detectors for five dimensions of affect: arousal, valence, expectation, intensity, and power on a continuous scale from -1 to +1, using Support Vector Regression (SVR) with polynomial kernels of degree 1 and radial base function (RBF) kernels as investigated in [48] for the two dimensional arousal and valence space. The LibSVM library [58] is used as SVR implementation.

For experimentation and evaluation we chose a maximum number of sessions (data from 49 sessions and 13 subjects) from the SEMAINE database [32] that have been annotated by the same coder (not all coders annotated all sessions). For subject-dependent cross-validation experiments we used all the data ($1,553$ user speech turns). Evaluation then has been done by conducting two-fold cross-validation over the full data set. For the subject-independent experiments we divided the data into two subsets: data from seven subjects vs. data from the other six subjects. Each subset was used once for training and once for testing. The measure of performance is the mean squared error (MSE) which is the mean of the sum of the squares of the prediction errors. MSE in each case is reported after averaging it over the two runs. Using these measures, we obtained the results presented in Table 3.

Additionally, a detector for interest is included. In the current system, this detector uses a Support Vector Machine (SVM) classifier with a polynomial kernel of degree 1 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. The recorded subjects were guided through an interactive commercial presentation, in order to elicit natural reactions with varying levels of interest. The database was annotated by four labelers. Since some components in the SEMAINE system require a continuous level of interest value, the centroid of the class confidences (the SVM probability estimates) is used. This rather unconventional approach has in practice shown better performance than a regression approach, which usually would be used for continuous outputs. Implementation of C++ on-line versions of the approaches presented in [59] and [60] will be the next promising step in this respect.

**Non-linguistic vocalisations.** For the automatic recognition of non-linguistic vocalisations we use Hidden Markov Models (HMM) trained on the SAL, SEMAINE, AMIDA, and AVIC databases. Currently, the following non-linguistic vocalisations are supported: breathing, coughing, hesitating, laughing, and sighing. In conformance with experiments and optimisations in [51], we apply left-to-right HMMs with nine states for every non-linguistic vocalisation. Even though the HMM topology used for the detection of non-linguistic vocalisations differs from the model topology applied for phoneme and keyword decoding, respectively, the non-linguistic vocalisations decoder is integrated into the keyword spotter component. Thus, the keyword spotting module outputs unified EMMA messages containing not only keywords but also non-linguistic vocalisations, together with their start times and confidences.

**Pitch contour analysis.** Based on the low level features probability of voicing and fundamental frequency (see Section 3.2.1) pseudo syllable units are identified. These units correspond to continuous voiced sections. Thereby the minimum unvoiced time between two voiced segments must be 20 ms, otherwise the two segments are treated as one. For each pseudo syllable segment, pitch direction statistics are computed using the differences between a long term average and a short term average moving window. Furthermore, the short term average at the beginning and end of each segment is compared. Based on this analysis, using a relative minimum variation threshold, a decision for five classes of pitch variation can be performed per syllable. These classes are: flat, rising, falling, rise-fall, fall-rise. Most important for the dialogue are the classes rise and fall. Since flat is usually the most frequent case, messages are only sent for the last four cases. The pitch direction information is sent as an EMMA message every time a pitch variation event is detected.

### *3.3.2 Affect analysis from visual features*

Affect analysis from visual features aims to achieve dimensional prediction of user affect (e.g., valence and arousal) from head gestures (amount and direction of head motion, and occurrences of head nods and shakes) and facial features on a continuous scale $[-1, +1]$. The current SAL system contains a head nod/shake detector, an affect predictor from head gestures, and an affect predictor from facial feature points.

**Head nod and shake detection.** Training data for head nod and shake detection was obtained by visually inspecting SAL [61] and SEMAINE [32] databases and manually cutting 100 head nod and 100 head shake clips of variable length. The directional codewords generated as part of the visual feature extraction module (Section 3.2) were fed into a Hidden Markov Model (HMM) for training a nod HMM and a shake HMM. However, to be able to distinguish other (somewhat noisy) head movements from the actual head nods/shakes, we (i) threshold the magnitude of the head motion, (ii) build an 'other HMM' to be able to recognise any movement but nods/shakes, and (iii) statically analyse the likelihoods outputted by the nod/shake/other HMM (maximum likelihood vs. training classifiers on the outputted likelihoods). In order to analyse the visual data continuously we empirically chose a window size of 0.6 secs that allows the detection of both brief and longer instances of head nods/shakes (as has been done by other related work, e.g. [62], [63]). From the global head motion features extracted and the head movements (nod or shake) detected, we created a window-based feature set (see [64] for details). The ground-truth for the window at hand consists of the dimensional annotations averaged over that window. Such a representation allows us to consider each feature vector independently of the others using the so-called static classifiers (i.e., regressors). We considered the Support Vector Machines for Regression (SVR) to the aim of dimensional emotion prediction from head gestures as they are among the most widely used regressors in the field [65]. Information about the head nod and head shake is sent as an EMMA message whenever an event is detected.

**Affect prediction from head gestures.** Dimensional emotion prediction from conversational head gestures aims to map the amount and direction of head motion, and occurrences of head nods and shakes onto arousal, expectation, intensity, power, and valence levels of the user. Results of affect prediction from head gestures are sent as an EMMA message with a frequency of 0.6 secs.

For experimental evaluation we used the same data (data from 49 sessions and 13 subjects from the SEMAINE database) and measures (2-fold cross validation, subject-dependent vs. subject-independent experiments) as for acoustic analysis (see Section 3.3.1). Taking into account the window size used for head nod and shake detection (0.6 secs), instead of speaker turns, we used segments of 30 video frames as units for prediction. Thus, our data set consists of $21,558$ instances obtained by processing $646,740$ video frames. The results of 5D affect prediction from head gestures are shown in Table 3. Further details are provided in [64]. Looking at the table, we conclude that (i) it is possible to predict arousal, expectation, intensity, power and valence dimensions from conversational head gestures and occurrences of nods and shakes, and (ii) dimensional emotion prediction from conversational head gestures supports generalisation across different subjects. Moreover, the table illustrates the fact that dimensional affect can be predicted from both types of cues (head gesture and voice) independently, and equally well. This finding should be treated with caution though. The suitability of MSE as an evaluation criterion for comparing the performance of dimensional affect recognisers remains to be tested. In future work, we will check the significance of this finding, and compare it to other evaluation measures such as correlation.

**Affect prediction from facial features.** Dimensional affect prediction from facial features aims to map the motion of facial feature points onto valence and arousal levels of the user (on a continuous scale $[-1, +1]$). The data for training the predictors were obtained from the SAL database [61]. The audiovisual SAL data from four subjects have been automatically segmented and appropriate annotations from multiple coders have been obtained (see [66] for details) for 134 segments. For prediction of arousal and valence (exhibited in every video frame), SVR were trained using the tracked positions of the facial feature points. Evaluation was based on subject-dependent 10-fold cross-validation (due to limited number of subjects at hand) using MSE (reported after averaging over the ten runs). Both polynomial (SVR-P) and radial-basis function (SVR-RBF) kernels were used for the experiments. For prediction of valence, best results for both SVR-P and SVR-RBF was MSE=0.054. For prediction of arousal, SVR-P provided slightly better results (MSE=0.088) compared to that of the SVR-RBF (MSE=0.090). Overall, we conclude that it is possible to predict arousal and valence dimensions from tracked facial feature points. The frequency of the EMMA message containing the affect analysis result depends on the predefined window size (set by modifying the config file). It is sent only if a face and its feature points have been detected continuously for the predefined window of video frames.

## 3.4 Dialogue management

Under 'dialogue management' we describe how *interpreters* and *action proposers* work together to determine *when* the agent speaks, what it says, and how it behaves while it is in the listener role.

The Dialogue Manager (DM) components are responsible for making sure that the conversation and interaction of the human with the virtual agent takes place. To do this, the dialogue manager needs to manage a number of things, such as superficial interpretation of the user behaviour, the turn taking behaviour, the backchanneling behaviour, and the utterance selection of the agent based mainly on the emotional state of the human.

The main challenge of these modules is to interpret some basic behaviours of the user and to provide natural feedback, without knowing much about the content. From the feature extractor and analyser components, the DM receives low-level features such as the energy of the audio, the F0 frequency, the position of the detected face, and facial points of that face. Analysers provide higher level features such as the arousal and interest of the user and some head gestures. From these features the emotional state of the user is the most important. Linguistic analysis is limited to crude keyword spotting. It is used to attempt to provide some coherence in the responses of the SAL characters. For instance, saying "Well done!" fits a context in which the user was telling the character about things the user did.

### *3.4.1 Speaking SAL*

As was explained in Section 1.2, the dialogues with the SAL agents 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 agents 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 reader to say more. The second criterion is determined by the SAL 'goal', i.e. to draw the interlocutor 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. Each module returns a list of possible responses - possibly empty - with for each response an estimate of the quality of that response (a value between zero and one). All these suggestions are grouped together. The quality of responses in the resulting set is lowered for those responses that have been used in recent turns. The response ending up with the highest value is then selected.

Currently, the following modules have been implemented.

- The 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.
- The Linking Sentence module suggests responses based on specified linking sentences. These are sentence pairs which can be linked by a typical user response. 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.".
- The Content Module suggests responses based on the detected keywords. This module is based on annotated transcriptions of the WOz recordings made in the Humaine project [31]. These transcriptions were annotated on certain high-level categories such as 'talking about past', 'talk about own feelings', 'agree/disagree', etc. On the basis of this data set a mapping was made between keywords and categories that is now being used for determining the categories with new input. A subset of the character utterances is tagged with the high-level categories. The Content module proposes responses based on the matching of the category set found through keyword spotting with the categories tagged on the character utterances.
- The 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.
- The 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?".

The selected verbal response is sent to the action selector component (see below).

Most of the models (all except the content-model) are based on common-sense of what a good response is, and on the input modules that are currently available. However, as data about this form of interaction between humans is becoming available (the Solid SAL data), we are moving to algorithms based on statistical analysis of these dialogues and what we can learn from machine learning. On top of the Solid SAL annotations, we have created another layer on this data, indicating which of the utterances of the SAL system would be a good response (impersonating the system) to the utterances of the user. Alternatively put, the layer replaces the utterances of the 'wizard' with utterances of the SAL system. Using plain statistical analysis and machine learning, we are learning which features of the user utterances can be used to select the respondent's utterance and thus what aspects of the user utterance the system should be sensitive to. Not only can this data and its analysis tell us more about the choices the system needs to make and for training the models, it can also be used for evaluation. Furthermore, it allows us to compare the performance on the annotated features with the performance on the information that is received from the input modules. Also the input modules can profit from the annotations and the analysis, as one can learn from the data which features are the most important and should be in the focus of the input modules (guaranteeing that what the system needs to be sensitive to is also sensed by the input modules).

The quality of the responses selected by the DM modules have been difficult to evaluate up to now in interactions, as the output of the modules depends heavily on the quality of the results provided by the input modules. As we just mentioned, having a growing corpus available for training the modules, we can start to make systematic evaluations of the contributions to the system that the dialogue manager has to offer. For instance, we can now compare the results of running the components using the input modules (keyword spotting and emotion recognition in particular) with running the components using the annotated corpus (the transcriptions and the set of emotion and other labels provided).

### *3.4.2 Listening SAL*

Whereas the speaker's communicative intentions are determined by the verbal planning modules described above, the listener's signals (called *backchannels* [67]) are automatically generated by the Listener Intent Planner module that is part of the *action proposers* components (see Fig. 4).

Studies have shown that there is a strong correlation between backchannel signals and the acoustic and visual behaviours performed by the speaker [68], [69]. From the literature [68], [69] 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. Then, 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) [6], [70] or signals of mimicry that mirror the speaker's signals.

The Action selection module [71] receives all the candidate actions coming from the action proposers. It has two roles. The first one is to manage the flow of candidate actions to be displayed by the agent. Indeed the Action Selection receives continuously candidate actions that are queued. 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 agents [72]. It also takes into account the emotions and interest level of the user, estimated through visual and acoustic analysis. Finally a candidate backchannel is chosen and sent to the behaviour generator.

We performed an evaluation of our backchannel selection module starting from two hypotheses which link Eysenck's two-dimensional representations of personality [72] with two variables: 1) backchannel frequency is hypothesised to be linked with the extroversion dimension and 2) backchannel type with the neuroticism dimension. Indeed (emotionally) unstable characters would perform less mimicry than (emotionally) stable ones [73], and extravert characters would perform more backchannels than introvert ones [74]. The evaluation study was performed via internet using a web browser showing videos of interactions between a user and a virtual agent. 93 participants (57 women, 37 men) judged for each video the frequency and the type of the backchannels according to the personality of the agent. The most recognised personality was aggressive (62%), followed by optimistic (53%), pessimistic (53%) and pragmatic (52%). The results show that the first hypothesis is partially verified: backchannel frequency can be viewed as an indicator of outgoing (t-test, $p < .01$) and pragmatic (t-test, $p < .05$) personalities. The second hypothesis is verified only for the pessimistic (Friedman test, $p < .001$) personality (i.e., Obadiah). These findings show that the selection of type and frequency of backchannels by the Listener Action Selection help to express some personalities.

### 3.5 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 the agent's communicative functions and some agent's behavioural characteristics (referred to as *baseline*). Its task consists in generating a list of behavioural signals for each communicative function. Each agent's *baseline* contains information on that agent's preference in using a given modality (speech, head, gaze, face, gesture, and torso) [75]. 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: e.g. wide vs. narrow gestures, fast vs. slow movements. All the possible communicative functions are associated with the multimodal signals that can be produced by the agent in order to convey them. Each of these associations represents one entry of the lexicon, called *backchannel lexicon*. Depending on the agent's baseline and the communicative function to convey, the system selects in the backchannel lexicon the most appropriate multimodal behavioural set to display. For example, an agent that wants to communicate its agreement could simply nod, or nod its head and smile, or say "m-hm".

The *audio synthesis* module uses MARY TTS [76] to synthesise both the spoken utterances and vocal backchannels like *myeah*, *uh-huh*, *oh*, etc. For the spoken utterances, we specially created expressive unit selection voices [77]. MARY TTS was extended to also allow for the generation of vocal backchannels [78]. For better lip synchronisation of audiovisual backchannels, our implementation consists of first generating the speech with timing information, using the same timing representation formats for text-to-speech and for listener vocalisations. All of the attributes of the vocalisation tag are optional; if an attribute is not given, this means that the search is not constrained on that level. The speech synthesiser looks up available vocalisations for the given speaker and generates the most appropriate vocalisation found for the request.

Finally, the multimodal behavioural signals are transformed into animation parameters following the MPEG-4 format [79], using Facial Action Parameters (FAP) and Body Action Parameters (BAP). Facial expressions, gaze, gestures and torso movements are described symbolically in repository files. Temporal information about the vocalisation, generated by the *audio synthesis* module, are used to compute and synchronise lips movements.

The animation is played in a graphic window by a FAP-BAP Player. Facial and body configurations are described through respectively FAP and BAP frames. The Player uses the OGRE graphics engine and DirectX9 technology to show one of the four SAL characters at a time.

We carried out a perception test in order to get a better understanding about multimodal backchannels and their interpretation by users. We asked subjects to judge a set of multimodal signals performed by the 3D agent Greta [80]. The signals were context-free, that is without knowing the discursive context of the speaker's speech. We hypothesised that the strongest attribution of a meaning will be conveyed by the multimodal signals obtained by the combination of visual and acoustic cues representative of the given meaning. 55 participants accessed anonymously to the evaluation study through a web browser. The multimodal signals were played one at a time. Subjects were asked to associate one or more meanings to each multimodal signal. We proposed twelve frequent meanings related to the listener's reactions during conversation [70], [81]: 'agreement', 'disagreement', 'acceptance', 'refusal', 'interest', 'no interest', 'belief', 'disbelief', 'understanding', 'no understanding', 'liking', 'disliking'. Participants used a bipolar 7-points Likert scale, where negative meanings were at one extreme and positive ones at the other: from -3 (extremely negative attribution) to +3 (extremely positive attribution). Using t-tests we found, for example, that the signal *nod+yeah* (N=12, mean=2.75) was more strongly judged as showing agreement than any other signal ($p<.05$), in line with our hypothesis. However, we also found that the signal *shake+no* (N=14, mean=-1.71) was not more strongly judged as showing disagreement than the other signals ($p>.05$), contradicting our hypothesis.

Results showed that the meaning conveyed by a multimodal backchannel cannot be simply inferred by the meaning of each visual and acoustic cues that compose it. It must be considered and studied as a whole to determine the meaning it transmits. Moreover, this evaluation allowed us to extend the backchannel lexicon by defining appropriate sets of acoustic and visual backchannel signals that the agent can display to convey a given communicative function.

## 4 Principles for evaluating Sensitive Artificial Listeners

It has been clearly recognised for some time that evaluating systems concerned with affect presents particular challenges [82]. Evaluating the SAL system brings together several of the challenges. This section deals with the principles of evaluation rather than the detailed findings.

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.

### 4.1 Principles for low-level evaluation

Low-level evaluations of the various components have been described above (see Section 3). 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 [83]: 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 [84], 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 [85]. 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 [36]. 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 [36]: 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 [86]; 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 [87]. 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.

### 4.2 Principles for high-level evaluation

Turning to evaluation of the system itself, the obvious starting point is the substantial literature on usability, which offers well-defined resources. However, it is 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." [88, 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. Clearly that does not cover everything that people might look for in an activity. As Edwardson put it, "We don't ski to be satisfied, we want exhilaration" [89, p. 2].

Westerman et al. [82] trace the development of more richly affective measurement systems. They document four main areas where measurement has developed: computer anxiety; trust and loyalty; frustration; and 'flow, fun, and playfulness'. The first two appear 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 them as people, then it is right and proper that we should be frustrated by Obadiah's relentless pessimism or Poppy's relentless brightness. Engagement is also a key issue in the last area. 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 (see [90]).

The point of these arguments is to draw out familiar types of test so that it is possible to see that they are not really relevant, and move past them to focus on the core issues. In essence, the key issue in this situation is whether users feel intuitively that they are engaged in a real conversation with a real personality; and linked to that, whether they respond realistically, despite knowing intellectually that the other party is not real. These are closely related to the issues that have been highlighted in research on presence (e.g. [91]). On that basis, we have developed and explored a collection of techniques that seem to be suited to this particular task. They involve three main elements.

1) **Verbal probes.** There are well-rehearsed reasons for being wary of verbal probes in the evaluation of affective devices (e.g. [92]). Asking for verbal reports during an emotional experience, particularly one that is also paradoxical, 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. Immediately following each interaction, a different character steps in and asks (orally) three questions about the interaction that has just finished. The questions target linked, but potentially separable aspects of the interaction:
   a) How naturally do you feel the conversation flowed?
   b) How often did you feel the avatar said things completely out of place?
   c) How much did you feel that you were involved in the conversation?

   The logic of the questions is that a) deals with the global structure of the exchange; b) deals with specific local anomalies within it; and c) deals with the user's involvement in the exchange, however orderly or otherwise it might be.
2) **Non-verbal concurrent task.** Users are given a button to hold during the conversation, and are asked to press it whenever they feel that the simulation is not working well. This provides a measure of engagement during the interaction. There is a degree of subtlety in it: the more engaged users are, the less likely they are to think of the button-press task, even if they do feel that the interaction is anomalous in some way.
3) **Objectively measurable signs of breakdown.** Interactions are recorded, and coded for behaviours that a combination of literature suggested may indicate a breakdown of engagement.

Results at this stage are incomplete, but promising. Interactions that are poorly rated on the verbal probes tend to show reduction in a range of behaviours, both visible (looking sideways or down, head movements, and hand gestures) and vocal (long utterances, amused laughs, exclamations); and increases in some vocal ones (nervous laughs, unfilled pauses, short utterances, sighing and audible breathing). Correlations among the items make a logical point worth noting. For two characters, Spike and Prudence, verbal and concurrent indicators intercorrelate in a single global evaluation; for the others, verbal probes a) and c) hang together, but b) and the concurrent task seem to be unrelated. The point is a simple one: what is anomalous for one character need not be anomalous for another. The implication is that measures need to be wary of assuming that 'one size fits all'.

These measures provide a basis for studying responses when the system is varied in a number of critical ways. In particular, we can vary the emotion detection modules that are active, and establish how engagement is affected by switching them on or off, or deliberately drawing the wrong conclusion from them. We can also study the effects of varying the conversation control strategies. The total outcome will be a systematic set of experiments which uses appropriate measures to assess the contribution that the components of the system make, singly or in combination, to the kind of interaction that it is designed to achieve.

## 5 Discussion and Conclusion

The previous sections have illustrated the complexity associated with the implementation of an autonomous SAL. There is a substantial divergence between the understanding from social and cognitive science how the phenomena involved in this type of system should be modelled, and our ability to implement them at the present time. Undoubtedly, our system is a grossly simplified version of what *should* be done; at the same time, it shows what *can* be done at present.

The integrated, autonomous SAL system as presented here is, first of all, a 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 items known within the system.

Despite its obvious limitations, thus, the current SAL system is a valuable starting point for research. The full system is publicly 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 parameterisations of the system are accessible such that, even without extensive 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.

Given its focus on the technical aspects and scientific principles, the present article touches only briefly on one essential aspect of creating an autonomous SAL: the evaluation of the system as a whole and of its parts. We have sketched what we think should be the principles for evaluating the system as a whole. Existing recipes for evaluating interactive systems do not transfer easily to a SAL scenario; even the question of the measurement tool itself is an open research question. Furthermore, we have pointed out properties and problems related to evaluating individual system components. A future publication will put these principles into practice and report on an evaluation of the SAL system and its parts.

## 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).

## References

[1] 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.

[2] 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.

[3] 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 *Language and Robots: Proceedings from the Symposium (LangRo'2007)*, 2007, p. 55–64.

[4] 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.

[5] 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.

[6] 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

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

[8] 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.

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

[10] M. Löckelt and N. Pfleger, "Multi-Party interaction with Self-Contained virtual characters," in *Proceedings of the 9th Workshop on the Semantics and Pragmatics of Dialogue (DIALOR)*, Nancy, France, 2005, p. 139–142. [Online]. Available: http://dialor05.loria.fr/Papers/21-Loeckelt.pdf

[11] P. Gebhard, M. Schröder, M. Charfuelan, C. Endres, M. Kipp, S. Pammi, M. Rumpler, and O. Türk, "IDEAS4Games: building expressive virtual characters for computer games," in *Proc. IVA*, vol. LNCS 5208. Tokyo, Japan: Springer, 2008, pp. 426–440. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-85483-8_43

[12] P. Gebhard, "ALMA - a layered model of affect," in *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-05)*, Utrecht, 2005.

[13] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "It's all in the game: Towards an affect sensitive and context aware game companion," in *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, Amsterdam, The Netherlands, 2009, pp. 29–35.

[14] K. O'Regan, "Emotion and e-learning," *Journal of Asynchronous learning networks*, vol. 7, no. 3, p. 78–92, 2003.

[15] R. Aylett, A. Paiva, J. Dias, L. Hall, and S. Woods, "Affective agents for education against bullying," in *Affective Information Processing*, J. Tao and T. Tan, Eds. London: Springer, 2009, pp. 75–90. [Online]. Available: http://dx.doi.org/10.1007/978-1-84800-306-4_5

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

[17] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1377280

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

[19] 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

[20] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual and spontaneous expressions," in *Proceedings of the 9th international conference on Multimodal interfaces*. Nagoya, Aichi, Japan: ACM, 2007, pp. 126–133. [Online]. Available: http://portal.acm.org/citation.cfm?id=1322192.1322216

[21] 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*, 2010.

[22] 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.

[23] 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

[24] 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.

[25] E. André and C. Pelachaud, "Interacting with embodied conversational agents," in *New trends in speech-based interactive systems*, F. Chen and K. Jokinen, Eds. New York: Springer, 2009.

[26] D. Heylen, E. Bevacqua, C. Pelachaud, I. Poggi, J. Gratch, and M. Schröder, "Generating listening behaviour," in *The HUMAINE Handbook on Emotion-Oriented Systems Technologies*, P. Petta, R. Cowie, and C. Pelachaud, Eds. Springer, 2010.

[27] 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

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

[29] 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: http://portal.acm.org/citation.cfm?id=365168&dl=GUIDE,

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

[31] 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.

[32] G. McKeown, M. Valstar, R. Cowie, and M. Pantic, "The semaine corpus of emotionally coloured character interactions," in *Proc. of the IEEE Conf. on Multimedia and Expo*, 2010, accepted.

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

[34] 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 *Proceedings of the ISCA Workshop on Speech and Emotion*, Northern Ireland, 2000, pp. 19–24.

[35] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of Long-Range dependencies," in *Proc. Interspeech*, Brisbane, Australia, 2008.

[36] 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.

[37] 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.

[38] E. Z. McClave, "Linguistic functions of head movements in the context of speech," *Journal of Pragmatics*, vol. 32, no. 7, pp. 855–878, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/B6VCW-40JG19M-1/2/5353b57cf44f092dd325164eb7dc750b

[39] 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: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1467-9280.2007.02024.x

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

[41] 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.

[42] A. S. Foundation, "Apache ActiveMQ," http://activemq.apache.org/. [Online]. Available: http://activemq.apache.org/

[43] M. Johnston, P. Baggia, D. C. Burnett, J. Carter, D. A. Dahl, G. McCobb, and D. Raggett, "EMMA: extensible MultiModal annotation markup language," Feb. 2009. [Online]. Available: http://www.w3.org/TR/emma/

[44] 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 *Intelligent Virtual Agents*, 2006, pp. 205–217. [Online]. Available: http://dx.doi.org/10.1007/11821830_17

[45] A. Berglund, S. Boag, D. Chamberlin, M. F. Fernández, M. Kay, J. Robie, and J. Siméon, "XML path language (XPath) 2.0," W3C Recommendation, Jan. 2007. [Online]. Available: http://www.w3.org/TR/2007/REC-xpath20-20070123/

[46] F. Eyben, M. Wöllmer, and B. Schuller, "openear - introducing the munich open-source emotion and affect recognition toolkit," in *Proceedings of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009)*, vol. I. IEEE, 2009, pp. 576–581.

[47] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - speech and music interpretation by large-space extraction, version 1.0.0," http://opensmile.sourceforge.net/, 2010. [Online]. Available: http://opensmile.sourceforge.net/

[48] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proceedings of 9th INTERSPEECH 2008*. Brisbane, Australia: ISCA, 2008, pp. 597–600.

[49] J. Carletta, S. Ashby, S. Bourban, M. Flynnand, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meetings corpus," in *Proceedings of the Measuring Behavior symposium on Annotating and measuring Meeting Behavior*, 2005.

[50] 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 (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, vol. 27, no. 12, pp. 1760–1774, 2009.

[51] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *Proc. 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-based Systems (PIT 2008), Kloster Irsee, Germany*, E. André, Ed., vol. LNCS 5078. Springer, 2008, pp. 99–110, 16.-18.06.2008.

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

[53] M. F. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 1–8, Mar 2010.

[54] H. Drucker, "Improving regressors using boosting techniques," *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pp. 107–115, 1997.

[55] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," *Proceedings of the Language Resources and Evaluation Conference*, pp. 1–6, Mar 2010.

[56] P. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[57] O. Jesorsky, K. Kirchberg, and R. Frischholz, "Robust face detection using the hausdorff distance," *LECTURE NOTES IN COMPUTER SCIENCE*, pp. 90–95, 2001.

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

[59] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 7–19, Mar. 2010.

[60] 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, Special Issue on Speech Processing for Natural Interaction with Intelligent Environments (to appear)*, 2010.

[61] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, L. Lowry, M. McRorie, L. Jean-Claude Martin, J.-C. Devillers, A. Abrilian, S. Batliner, A. Noam, and K. Karpouzis, "The humaine database: addressing the needs of the affective computing community," in *Proc. of the Second Int. Conf. on Affective Computing and Intelligent Interaction*, 2007, pp. 488–500.

[62] A. Kapoor and R. W. Picard, "A real-time head nod and shake detector," in *Workshop on Perceptive User Interfaces*, 2001.

[63] W. Tan and G. Rong, "A real-time head nod and shake detector using hmms," *Expert Systems with Applications*, vol. 25, no. 3, pp. 461–466, 2003.

[64] H. Gunes and M. Pantic, "Dimensional emotion recognition from spontaneous head gestures for interaction with sensitive artificial listeners," in *Proc. Int. Conf. on Intelligent Virtual Agents*, 2010, submitted.

[65] ——, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

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

[67] V. Yngve, "On getting a word in edgewise," in *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, 1970, pp. 567–577.

[68] 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.

[69] 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.

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

[71] E. de Sevin and C. Pelachaud, "Real-time backchannel selection for ecas according to user's level of interest," in *International Conference on Intelligent Virtual Agents*, Amsterdam, Holland, September 2009.

[72] E. Bevacqua, E. de Sevin, C. Pelachaud, M. McRorie, and I. Sneddon, "Building credible agents: Behaviour influenced by personality and emotional traits," in *International Conference on Kansei Engineering and Emotion Research (KEER'10)*, Paris, France, 2010.

[73] P. Borkenau and A. Liebler, "Trait inferences: Sources of validity at zero acquaintance," *Journal of personality and social psychology*, vol. 62, no. 4, pp. 645–657, 04 1992.

[74] T. Chartrand and J. Bargh, "The Chameleon Effect: The Perception-Behavior Link and Social Interaction," *Personality and Social Psychology*, vol. 76, pp. 893–910, 1999.

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

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

[77] S. Pammi, M. Charfuelan, and M. Schröder, "Multilingual voice creation toolkit for the MARY TTS platform," in *Proc. LREC 2010*, Malta, 2010.

[78] S. Pammi, M. Schröder, M. Charfuelan, O. Türk, and I. Steiner, "Synthesis of listener vocalisations with imposed intonation contours," in *7th ISCA Speech Synthesis Workshop*, Kyoto, Japan, 2010, submitted.

[79] J. Ostermann, "Face animation in MPEG-4," in *MPEG-4 Facial Animation - The Standard Implementation and Applications*, I. Pandzic and R. Forchheimer, Eds. Wiley, England, 2002, pp. 17–55.

[80] 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.

[81] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Semantics*, vol. 9, no. 1, 1993.

[82] 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

[83] 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

[84] 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.

[85] 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, Aug. 2005. [Online]. Available: http://psr.sagepub.com/cgi/content/abstract/9/3/184

[86] M. Brendel, R. Zaccarelli, B. Schuller, and L. Devillers, "Towards measuring similarity between emotional corpora," in *Proc. LREC workshop on Emotion Corpora*, Valettea, Malta, 2010, pp. 58–64.

[87] J. Bachorowski, "Vocal expression and perception of emotion," *Current Directions in Psychological Science*, vol. 8, no. 2, pp. 53–57, 1999.

[88] I. I. O. for Standardization, "Ergonomic requirements for office work with visual display terminals (VDTs) - part 11: Guidance on usability," International Standards Organisation, ISO Standard ISO 9241-11, 1998.

[89] M. Edwardson, "Measuring consumer emotions in service encounters: an exploratory analysis," *Australasian Journal of Market Research*, vol. 6, no. 2, p. 34–48, 1998.

[90] M. Csikszentmihalyi and I. S. Csikszentmihalyi, *Beyond boredom and anxiety*. San Francisco, USA: Jossey-Bass, 1975.

[91] 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

[92] K. Isbister, K. Höök, M. Sharp, and J. Laaksolahti, "The sensual evaluation instrument: developing an affective evaluation tool," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. Montréal, Québec, Canada: ACM, 2006, pp. 1163–1172. [Online]. Available: http://portal.acm.org/citation.cfm?id=1124772.1124946