Towards responsive Sensitive Artificial Listeners

Marc Schröder[1], Roddy Cowie[2], Dirk Heylen[3], Maja Pantic[4], Catherine Pelachaud[5], Björn Schuller[6]

Abstract. This paper describes work in the recently started project SEMAINE, which aims to build a set of Sensitive Artificial Listeners – conversational agents designed to sustain an interaction with a human user despite limited verbal skills, through robust recognition and generation of non-verbal behaviour in real-time, both when the agent is speaking and listening. We report on data collection and on the design of a system architecture in view of real-time responsiveness.

1 INTRODUCTION

Conversations between humans and machines today are substantially different from conversations between humans. While humans can talk to one another for sustained periods, possibly hours, and may give limited importance to the actual content of the interaction, human-machine dialogue is often task-oriented and finishes as soon as the task is fulfilled.

One major reason for the inability of current dialogue systems, including multimodal ones, to sustain an interaction with the user is their inability to provide truly responsive reactions, including non-verbal behaviour. Specifically, current dialogue systems do not appear to be *listening* when in the listener role – they are usually silent and, if they have a visual appearance, may perform some idle behaviour which however is unrelated to the user's utterance.

The FP7 project SEMAINE (1/08-12/10) will build a multimodal dialogue system with an emphasis on non-verbal skills – detecting and emitting vocal and facial signs related to the interaction, such as backchannel signals expressing continued presence, attention or interest, an evaluation of the content, or an emotional connotation. The system to be built has strong real-time constraints, because it must react to the user's behaviour while the user is still speaking.

The project reflects a strategy of converging streams. One of the streams already exists. It consists of systems which have complex linguistic skills, but lack the non-verbal stream that humans use. There is a tendency to assume that the way to move forward from that must be to graft non-linguistic skills gradually onto the linguistic structures. However, reflection raises several doubts about that route. It is certainly not the way biological communication operates. There non-verbal communication is primary, and verbal communication is grafted onto it (both ontogenetically and phylogenetically). The route also runs into deep difficulties because speech recognition is still weak. Natural non-verbal communication presupposes speech as the medium for transmitting language, but it is very challenging to extract the words needed for complex linguistic processing from the kind of communication that is rich in non-verbal signals.

The SEMAINE strategy is to complement the linguistic stream by building systems that have rich non-verbal skills, but very little truly linguistic competence. It seems that the goal should be achievable in principle. There is ample informal evidence that humans can sustain that kind of interaction – for instance, at a party where there is too much noise to make out many of the words that someone is saying; or in interactions between people who want to convey goodwill, but do not speak each other's languages. Once that non-verbal stream is established, it becomes possible to explore alternative approaches to integrating the two.

SEMAINE explores a particular kind of system, which has been called a Sensitive Artificial Listeners (SAL). SAL systems are like the well-known Eliza chatbot [1] in that they use very shallow verbal skills in ways that users find it easy to engage with. Unlike Eliza, they rely on non-verbal signals from the user to guide selection of responses from their limited repertoire. Previous Wizard-of-Oz studies [2][3] have shown that the scenario can be developed in ways that engage users in quite sustained interactions. One of the keys to success is that the system's utterances are designed to elicit emotion. That gives rise to relatively strong non-verbal signals, which can be used to choose at least roughly appropriate reactions.

The present paper describes early work in the project: data collection, used to provide evidence for human behaviour, and system integration in view of real-time responses.

2 DATA COLLECTION

The particular kind of SAL that SEMAINE uses allows a user to interact with one of four characters (known as SAL agents). Each agent has a particular emotional agenda. Obadiah tries to make users sad; Spike tries to make them angry; Poppy tries to make them happy; and Prudence tries to make them reasonable. Users are free to choose which agent they will talk to at any given time. The basic definition of each agent's character is given by a script listing responses that the agent may give if the user is sad, responses it may give if the user is angry, and so on. The scripts evolved during previous projects, and are known to support reasonably sustained interactions.

SEMAINE depends in multiple ways on data from recordings of users interacting with SAL. The most traditional issue is recognising the user's emotional state in order to decide which set of responses it is appropriate to use. It is widely accepted that recognition will be poor unless recognisers are trained on data collected in a situation that mirrors the use scenario closely [4].

A less standard set of issues involves detailed description of the non-verbal behaviour used to sustain and direct conversations (such as nods, eye movements, non-verbal sounds, etc). Pre-existing knowledge about these issues is limited. There are

[1] DFKI GmbH, Saarbrücken, Germany. Email: schroed@dfki.de.
[2] Queen's University, Belfast, Northern Ireland.
[3] HMI, Universiteit Twente, Netherlands.
[4] Imperial College, London, UK.
[5] Université Paris 8 / INRIA, Paris, France.
[6] Technische Universität München, Germany.

general statements, many of them long established. However, neither the vocal, facial or bodily *forms* nor their possible *functions* are documented in anything approaching the level of detail that would be necessary to recognise and generate non-verbal signals in the context of an interaction with any reasonable degree of accuracy.

Last, and in some ways most challenging, data is needed to derive rules to specify how agents should select within the set of options consistent with the user's current emotional state. Pilot studies indicate that human operators do not need to know exactly what users have said in order to avoid obviously anomalous choices; non-verbal cues carry most of the relevant information. Very little is known about the form the cues take, or the dimensions onto which they map.

For these reasons, SEMAINE depends heavily on data. It is provided by a graded set of scenarios, which simulate more and more closely the kind of interaction to be modelled.

The coarsest level of simulation already exists. The SAL agents were simulated by a human operator who selected what she judged to be the appropriate response from a SAL script and read it with the appropriate tone of voice [5]. Audio-visual recordings of the users show rich expressive behaviours; annotated selections are available in the HUMAINE database [2][6].

The second level of simulation will use a scenario called Solid SAL. Its core function is to provide information that is completely missing from the initial SAL recordings; that is, information about the behaviour that the SAL character should display while the user is speaking. To obtain that, human actors will impersonate the various SAL characters.

The recordings have to strike several balances. The communication will be mediated through cameras, microphones and computer screens, in order to give the user an experience that is as close as possible to interaction with a machine. Similarly, the actors will adhere to the spirit of the scripts that define the computerised characters. However, the interaction needs to have a high level of spontaneity to provide the key information. Hence, for instance, the actors must not be looking at a script to find suitable responses when normal interaction would require them to be looking at the user.

The last scenario, Wizard-of-Oz SAL, is closer again to actual human-machine interaction. The human user believes he/she is interacting with a computer program; however, in reality, pre-recorded vocal utterances for the SAL characters are chosen by a human Wizard operator in a different room, who can see and hear the human user. This setup will not give data on appropriate agent behaviour, but it will produce data that simulates specific problems related to human-machine interaction, such as confusion with inappropriate machine answers.

We aim to record a total of 20 hours of SAL interactions in two main simulation scenarios. Key parts of the task will be to provide recordings in a format that facilitates machine analysis, and to develop and implement a suitable labelling.

Recording format is a major issue for automatic interpretation. It is one of the attractions of the SAL scenario that it avoids some of the problems that tend to accompany material with any degree of naturalism. It is natural for the user to sit in front of a screen, with relatively limited movement of the head or whole body. As a result, images of the face are usually close to full frontal. Because communication is apparently with a computer, microphones and cameras are integral to the scenario, and do not distract from the interaction; and they can be positioned to optimise recording rather than having to be unobtrusive.

Equipment is another issue. The SAL recordings in the HUMAINE database used home quality cameras and microphones. That poses problems for video analysis in particular. High spatial resolution is needed to identify key points associated with, for instance, the corners of the mouth or the direction of gaze. Significant gestures can also be very rapid, and that requires high temporal resolution. Some informative movements have a significant forward-to back component, and that is difficult to recover from a single frontal camera. The recording setup has been designed to address these issues.

The user's speech is recorded using an AKG HC 577 head microphone and an AKG C 1000 S directional microphone. Frontal views are provided by a Stingray F-046C camera (colour) and a Stingray F-046B (mono, for high spatial resolution), both giving 780*580 pixels and 61fps; and a second F-046B gives a side view. In the Solid SAL scenario, the operator (in a separate room) has a similar recording configuration, but without the side view camera. The StreamPix4 program is used to control the recording of the five cameras, and record directly to high capacity discs.

The major challenge in annotating the recordings is to identify a manageable subset of the potentially interesting behaviours. Three main types are currently envisaged.

High-level states. These will be recorded using the trace techniques of the type pioneered in the HUMAINE database [2][6]. The classical dimensions of affect (valence, activation, potency) are a natural starting point. Of the 'basic emotions', happiness, surprise and anger will be annotated: the others are likely to be rare or absent. Several 'epistemic / affective' states of the type highlighted by Baron-Cohen [7] are likely to be important, notably agreement, and interest. Trace-type techniques also seem appropriate to capturing high-level attributes of linguistic expression, such as emphasis and questioning tone.

Linguistic and paralinguistic. Spoken content will be transcribed, with standard punctuation. Beginnings and ends of turns will be recorded, as will those of long pauses. Vocalisations such as laughing, yawning, audible breathing, and coughing will be annotated, with laughter being divided into categories such as voiced/unvoiced, melodic, positive/negative.

Visible behaviour. Annotation of facial action units is a basic requirement to train expression recognition. Head actions such as nodding will be annotated, and distinguished according to form (e.g. sustained, repeated) and function (e.g. 'continue', 'agree'). A coarse description of gaze direction will also be annotated.

The resulting material will be made generally available. It will be a major resource for research in general, not just SEMAINE. In the past, it has been been difficult to access conversation that is natural, emotionally coloured, recorded to high specifications, and annotated in some depth. The labelling exercise should also be a substantial contribution in its own right, since it is not clear at present what level of labelling is needed to support broadly acceptable non-verbal behaviour. SEMAINE aims to identify empirically a minimum set of descriptors needed for a functional system.

3 TOWARDS A REAL-TIME SYSTEM

The creation of an autonomous SAL system depends on a suitable architecture.

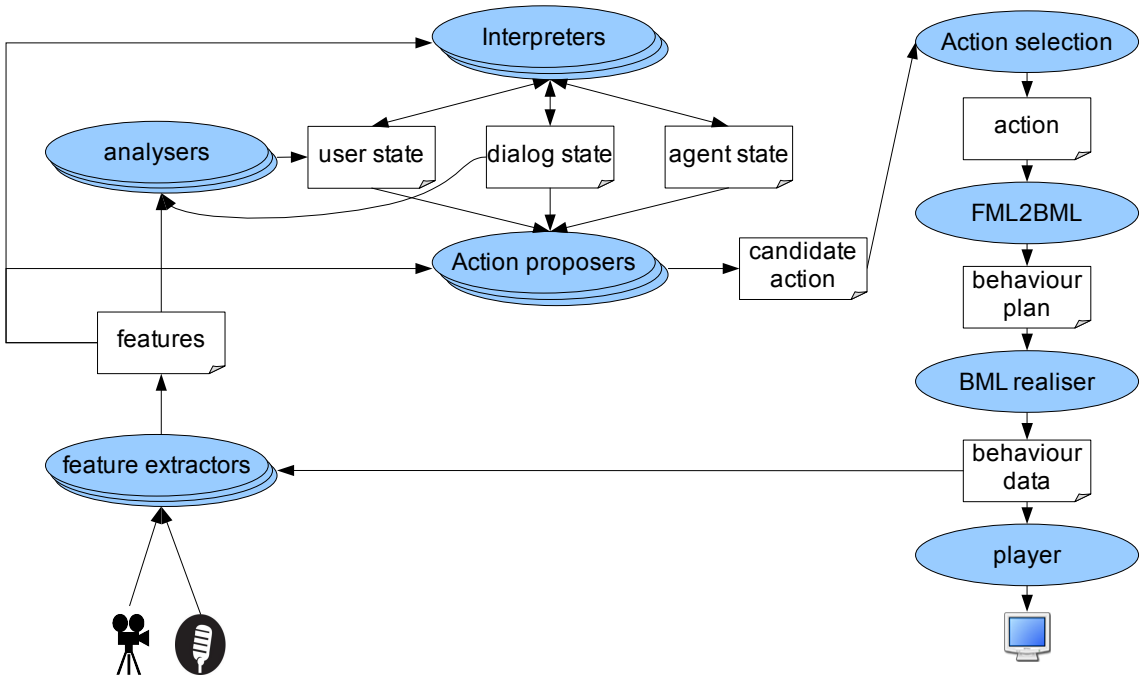

Figure 1: Architecture of the initial SEMAINE system

3.1 A "pipeline" system architecture

Figure 1 shows a first approximation of our concept. Processing components are represented as ovals, data as rectangles. Arrows are always between components and data, and indicate which data is produced by or is accessible to which component.

It can be seen that the rough organisation follows the simple tripartition of input (left), central processing (middle), and output (right), and that arrows indicate a rough pipeline for the data flow, from input analysis via central processing to output generation. Note that this is a deliberate simplification at this stage, and is identified as a point for future improvement – see Section 3.5.

The main aspects of the architecture are outlined as follows. Feature extractors analyse the low-level audio and video signals, and provide feature vectors periodically to the following components. A collection of analysers, such as monomodal or multimodal classifiers, produce a context-free, short-term interpretation of the current user state, in terms of behaviour (e.g., a smile) or of epistemic-affective states (emotion, interest, etc.). These analysers usually have no access to centrally held information about the state of the user, the agent, and the dialog; only the speech recognition needs to know about the dialog state, whether the user or the agent is currently speaking.

A set of interpreter components evaluate the short-term analyses of user state in the context of the current state of information regarding the user, the dialog, and the agent itself, and update these information states.

A range of action proposers produce candidate actions, independently from one another. An utterance producer will propose the agent's next verbal utterance, given the dialog history, the user's emotion, the topic under discussion, and the agent's own emotion. An automatic backchannel generator identifies suitable points in time to emit a backchannel. A mimicry component will propose to imitate, to some extent, the user's low-level behaviour. Finally, a non-verbal behaviour component needs to generate some "background" behaviour continuously, especially when the agent is listening but also when it is speaking.

The actions proposed may be contradictory, and thus must be filtered by an action selection component. A selected action is converted from a description in terms of its functions into a behaviour plan, which is then realised in terms of low-level data that can be used directly by a player.

Similar to an *efferent copy* in human motor prediction [8], behaviour data is also available to feature extractors as a prediction of expected perception. For example, this can be used to filter out the agent's speech from the microphone signal.

3.2 System integration

In order to build an integrated system out of partly existing components, SEMAINE has adopted an early integration paradigm: Before even starting to tune system components to the SAL domain, we build a first integrated instance of the system within the first project year. The concrete process of establishing system-level communication between processing components serves as a "reality check" for conceptual modelling: any conceptual unclarities become apparent and can be addressed early in the project. This means that the first integrated system will neither be real-time nor implement the functionality of a SAL, but it lowers the barrier for both by providing a basis for domain modelling and system optimisation.

Integrating existing processing components into one system, across different programming languages and operating systems, requires a middleware. We have selected a message-oriented middleware, ActiveMQ [9], a fast, open-source implementation of the Java Message Service (JMS) with clients available in a range of programming languages.

To speed up the integration, we have implemented a dummy version of the pipeline system architecture: all components exist,

data in the right kind of representation format is flowing between them, but none of the components does anything meaningful yet. The dummy system can be thought of as a code-level formulation of the architecture diagram in Figure 1. It has two important benefits:

- it makes it easier to integrate the actual components, because any given component can be tested in the context of the dummy system, by replacing the dummy component with the "real" one (see Section 3.3);
- it must find at least provisional answers to basic integration issues such as the representation formats for the various kinds of data, and thus highlights the need to define these carefully (see Section 3.4).

The following section gives an overview of the technology that will be introduced in the system.

3.3 System components

The SAL system will observe the user's facial expression, gaze, and voice, and pick up a word from time to time; it will react through a combination of low-level feedback mechanisms and planned dialogue contributions, which will be realised through a synthetic voice and an embodied conversational agent (ECA).

Preliminary versions of most components exist with the project partners; the challenge at this stage is to integrate them into a suitable architecture, to ensure the data flow, and to set up mechanisms for real-time reactivity.

3.1.1 Input components

Concerning speech analysis, SEMAINE will use one centralised Low-Level-Descriptor extraction module providing information on pitch, intensity, durations, formants, harmonics-to-noise-ratios, cepstral and further spectral contours, etc. As mid-level feature further serves the spoken content by a speech recogniser tailored for spontaneous affective speech [10]. A further key-word spotter focuses on a limited vocabulary of interest for dialog control. We aim to recognise states such as emotion [11] and interest [12] through incremental processing [13], but also non-linguistic vocalisations [14] such as laughter, consent or hesitations, and sentence modality. These analyses rely on a combined brute-force acoustic feature generation based on hierarchical statistical functionals. Likewise borrowing from the same feature pool of several thousands, cost-sensitive feature space optimisation is carried out once per task. Diverse machine learning algorithms such as Hidden Conditional Random Fields, Dynamic Bayesian Networks or Long-Short-Term-Memory recurrent networks serve dynamic and static modelling needs and care for task-specific requirements as history integration or asynchronous fusion. Main challenges will be adaptation and user modelling, incremental processing, and utmost robustness in view of highly spontaneous speech and ambient noises [10][11].

The vision part of SEMAINE includes simultaneous face tracking, head pose estimation, and facial feature points tracking (including irises), in order to address the problem of detecting and tracking the face and its features in naturalistic settings, regardless of head pose, clutter, and variations in lighting conditions. A particle-filtering tracking scheme will be used to achieve this [15][16]. The target is to devise an all-in-one approach to facial gesture recognition including facial muscle actions (AUs, [17][18]), visual focus of attention, and head nods, such that it is informed by head pose and facial points tracker.

SEMAINE also includes audiovisual analysis of user behaviour. On lower semantic levels this relates to the analysis of behavioural cues such as affect bursts, which include laughter, sighs like 'uffff' and 'ahhh', and disfluencies. Spoting these vocalizations is based on the related auditory profiles, related facial movements like smiles, and related head movements [19][20].

On a higher semantic level, the target is to develop a set of audiovisual methods for detecting human affective states including the user's positive and negative reactions to a SAL character and his or her attitude like agreeing and disagreeing. Detection will be done based on morphological and temporal correlations between relevant behavioural cues, including facial gestures like frowns and smiles [21], head gestures like tilts and nods, visual focus of attention (gaze direction), acoustic and linguistic feature information, and vocal outbursts like laughter and sighs. Main challenges in this domain include attaining suitable temporal reasoning about correlations across different modalities and across different behavioural cues as well as addressing the open issue related to the level at which multimodal data fusion should be achieved (feature level, mid-parameter level, or decision level) [22][23].

3.1.2 Action planning components

As conversational agents, the SAL characters need to enter into a collaborative enterprise, a conversation, with a human interlocutor. The contributions they make to the dialog should be *appropriate* at each point in time. Also the contributions are continuous. If the characters stop speaking, they continue to deliver contributions through facial expressions, head movements and eye gaze. Contributions are thus multimodal including verbal, nonverbal and paraverbal elements.

In general terms one can think of the characters as being governed by the Gricean maxims of conversation, relevance in particular, in the same way as humans are in conversations. In order for contributions to be relevant they should take into account various dimensions of the context (history of interaction, setting, goals of the conversation...). For instance, the topic of the conversation cannot be changed at random; characters should follow the conventional procedures of how to change topic ("by the way"...). In the case of SAL characters it is, for example, not always appropriate to ask the interlocutors whether they want to talk to another character. Also important is the fact that the contributions should be made at the appropriate time.

Besides the natural language utterances, that are more or less prefixed in the SAL system, the contributions consist of verbal and nonverbal backchannels. Also, the face and head should provide a running commentary to what the speaker is saying, which may involve mimicry on the side of the agent, and they should complement the utterance of the SAL character as it is speaking itself. The selection of a contribution (backchannel and utterance) will not involve intricate planning nor will the system engage in deep semantic analysis of the utterances by the speaker, but there are several parameters that need to be taken into account in the selection of all the contributions. For the utterances this involves some aspects of content. For instance, some SAL utterances ("Well done" would be such an example) are only appropriate if the speakers talked about something they did themselves. Also many backchannels are sensitive to some aspects of content. Furthermore, on the structural level, the action selection should take into account notions such as turn-yielding signals or backchannel elicitors (involved in grounding processes).

Action selection depends critically on how fast and accurate the various classifiers and interpreters can compute the relevant information. In accordance with its personality settings, the character needs to make a decision on its communicative intentions (show engagement by emitting a backchannel or reacting negatively by an utterance).

3.1.3 Output components

Actions are described either directly in terms of behaviour (e.g., for mimicry), or in terms of abstract communicative functions. In order to generate system behaviour for these actions, a function-to-behaviour component (FML2BML in Figure 1) must identify suitable behaviour. A multi-modal *gestuary* contains the available behaviour, in one or several modalities such as face, gesture, or vocalisations, along with functional interpretations. In order to realise a certain communicative function, one of the associated behaviours is selected, taking into account any pre-existing allocations of modalities.

Any time alignments in the behaviour markup produced at this stage are symbolic, realised by cross-referencing between different elements of the markup. This is inevitable because concrete timing is not yet available at this stage.

Realisation of behaviour is performed in two stages. The speech synthesizer MARY [24] is used to convert text or speech synthesis markup into audio data, and to enrich the behaviour markup with detailed timing information.

The realisation of visual behaviour is performed by a component of the Greta system [25], which converts the symbolic but time-aligned description of behaviour into a low-level parametric description in terms of MPEG-4 Facial Action Parameters (FAPs) and Body Action Parameters (BAPs).

Finally, the low-level behaviour data is realised as an audio-visual presentation by the Greta rendering engine [25], interpreting the MPEG-4 action parameters with a concrete facial and body model, and displaying the visual behaviour in synchrony with the audio playback.

3.4 Data representation

How to represent the data flowing between components is a relevant research question in its own right. In view of future reuse, we consider it important to use standard data formats at clearly defined component interfaces wherever possible.

It appears that standards (or pre-standards prepared by dedicated initiatives) are more readily available on the output side than on the input side. SSML [26] for speech synthesis, MPEG-4 FAP and BAP parameters [27] for facial and body animation, are actual standards from recognised standardisation bodies. BML, a behaviour markup language for ECAs [28], is approaching a reasonable degree of maturity. FML [29], the counterpart to BML for functional representations, is at an earlier stage of development; here, the concrete implementation in SEMAINE may actually support the development. The specific kind of function that is related to emotional connotation can be represented using the emotion markup language EmotionML [30] currently under preparation at the W3C.

On the input side, low-level features such as pitch values and facial action units may best be represented as simple feature vectors. Immediate analyses of these, as produced by the analysers in Figure 1, can be embedded into W3C's extensible multi-modal annotation language, EMMA [31]. However, EMMA is more of a container for concrete annotations than an annotation format as such, so that representations are needed for the analysis outcomes. It may be possible to represent epistemic-affective states using EmotionML, thus yielding a symmetry between input and output regarding the representation of emotions and related states. We will evaluate whether the same symmetry can be attained on the level of behaviour, i.e. whether BML can be used within EMMA containers to represent analyses of user behaviour. Given that BML was developed for driving ECA output, it is not immediately clear if the representations available in BML are appropriate for representing input.

The information regarding the system's internal representation of current user state, dialog state, and agent state, is the most domain-specific aspect of the system. Here, a custom representation format is required in order to provide the concepts that play a role in the SEMAINE system.

3.5 Real-time architecture

A system showing a reactivity similar to that of humans has obvious strong real-time requirements. The system must ensure that feedback behaviour can be executed at the right moment. For example, Ward and Tsukahara [32] suggest that, for Japanese speakers, a backchannel should be executed 350 ms after a low-pitch section of a certain minimum duration in the user's speech. If the backchannel is produced earlier or later, it is perceived as less appropriate. The SEMAINE system must make sure that short and, ideally, well-defined reaction times can be achieved.

On the one hand, this depends on efficient processing – for example, redundancy should be avoided, e.g. by performing signal enhancement once and making the result available to all analysers that require it. Similarly, computationally intensive tasks should run on dedicated hardware machines, so that they do not slow down one another.

At least as importantly, however, is the architecture itself. The pipeline architecture as depicted in Figure 1 is very inefficient in the sense that the reaction time is the sum of the processing times of all components. For example, in order to emit a backchannel according to the rules proposed by Ward and Tsukahara [32], at least the following components would need to be executed: pitch assessment analyser, backchannel action proposer, function-to-behaviour mapping, text-to-speech, and visual planning to generate facial movement parameters in synchrony with the audio. In the example of 350 ms response time, that leaves an average 70 ms processing time per component – even with highly efficient components, this value will be difficult to guarantee. Consequently, a pipeline architecture will nearly inevitably show delays in responding.

In order to alleviate this problem, we intend to revise our system architecture once a basic system is up and running. Current plans for a more efficient architecture are inspired from cognitive science. Studies of action planning and action selection from cognitive neuroscience [33] are proposing models of degree-of-activation of *several* alternative actions *in parallel*; at a given time, the one action with the highest activation is selected and executed. Adapted to the SEMAINE architecture, this means that several potentially needed actions can be pre-generated, e.g. based on frequency of occurrence or based on early predictors. Such action candidates are prepared by the speech synthesis and behaviour generation components up to the point where the input data for the actual player is available. Only at the very last moment, a trigger component decides which of the available actions is most suitable; after selection, it can be immediately rendered.

Further improvements can be considered on the input side, by allowing the possibility to "re-consider" crucial input. Concretely, an interpreter component (in the sense of Figure 1) may identify a certain part of a recent user utterance as a crucial event and thus request a more fine-grained analysis from feature extractors or analysers for that period of time.

4 DISCUSSION

The work described in the present paper is at a very preliminary stage – we have only started to build an integrated system, and are still collecting the "ground truth" data on which the domain logic of the Sensitive Artificial Listeners will be modelled. The ambitious aim is a system exhibiting the kinds of non-verbal behaviour required so that a human user "feels like" he/she is actually communicating with the ECA.

Building a system such as ours is a promising way of doing research: far from being a mere engineering challenge, it forces us to bring abstract considerations into a concrete form, and to seek for new paradigms to make processing efficient and robust.

The SEMAINE data and system will also be made available to the research community. Large parts, including the system integration API, will be available under an open source license; some individual components will be available in binary form only and under a research license. This approach, in combination with the extensive use of standards, will make it possible to continue to use and extend the SEMAINE system and its components well beyond the limits of the project duration.

In particular, in the future, it will be interesting to investigate how traditional dialogue plans can be combined with the non-verbal competence of the SEMAINE system, thus ultimately leading to systems that do both: they feel natural to interact with, and they do something useful.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211486 (SEMAINE).

REFERENCES

[1] J. Weizenbaum, "ELIZA - a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, 1966, pp. 36-45.

[2] E. Douglas-Cowie et al., "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data," *Affective Computing and Intelligent Interaction*, 2007, pp. 488-500.

[3] D. Heylen, A. Nijholt, and M. Poel, "Generating Nonverbal Signals for a Sensitive Artificial Listener," *Verbal and Nonverbal Communication Behaviours*, 2007, pp. 264-274.

[4] A. Batliner et al., "How to find trouble in communication," *Speech Communication*, vol. 40, 2003, pp. 117-143.

[5] E. Douglas-Cowie et al., "The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation," Marrakech, Morocco: 2008, pp. 1-4.

[6] "HUMAINE Database download"; http://emotion-research.net/download/pilot-db/.

[7] S. Baron-Cohen et al., *Mind Reading: The Interactive Guide to Emotions*, London: Jessica Kingsley Publishers, 2004.

[8] D.M. Wolpert and J.R. Flanagan, "Motor prediction," *Current Biology*, vol. 11, 2001, pp. R729-R732.

[9] "Apache ActiveMQ"; http://activemq.apache.org/.

[10] B. Schuller, J. Stadermann, and G. Rigoll, "Affect-Robust Speech Recognition by Dynamic Emotional Adaptation," *Proc. Speech Prosody 2006*, Dresden, Germany: 2006.

[11] B. Schuller et al., "Towards More Reality in the Recognition of Emotional Speech," *Proc. ICASSP*, 2007, pp. IV-941-IV-944.

[12] B. Schuller et al., "Audiovisual recognition of spontaneous interest within conversations," *Proc. ICMI*, Nagoya, Japan: 2007, 30-37.

[13] M. Wöllmer et al., "Abandoning Emotion Classes - Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies," *Proc. Interspeech*, Brisbane, Australia: 2008.

[14] B. Schuller, F. Eyben, and G. Rigoll, "Static and Dynamic Modelling for the Recognition of Non-verbal Vocalisations in Conversational Speech," *Perception in Multimodal Dialogue Systems*, 2008, pp. 99-110.

[15] I. Patras and M. Pantic, "Particle filtering with factorized likelihoods for tracking facial features," *Proc. Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004.*, Seoul, Korea: 2004, pp. 97-102.

[16] I. Patras and M. Pantic, "Tracking deformable motion," *Proc. IEEE International Conference on Systems, Man and Cybernetics, 2005*, Waikoloa, Hawaii: 2005, pp. 1066-1071.

[17] M. Valstar and M. Pantic, "Fully Automatic Facial Action Unit Detection and Temporal Analysis," *Proc. CVPRW'06* , New York, USA: 2006, p. 149.

[18] M. Valstar and M. Pantic, "Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics," *Human–Computer Interaction*, 2007, pp. 118-127.

[19] S. Petridis and M. Pantic, "Audiovisual discrimination between laughter and speech," *Proc. ICASSP*, Las Vegas: 2008, 5117-5120.

[20] S. Petridis and M. Pantic, "Fusion of audio and visual cues for laughter detection," *Proceedings of the 2008 international conference on Content-based image and video retrieval*, Niagara Falls, Canada: ACM, 2008, pp. 329-338.

[21] M.F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," *Proc. ICMI*, Nagoya, Aichi, Japan: 2007, pp. 38-45.

[22] Z. Zeng et al., "A survey of affect recognition methods: audio, visual and spontaneous expressions," *Proc. ICMI*, Nagoya, Aichi, Japan: 2007, pp. 126-133.

[23] M. Pantic et al., "Human-Centred Intelligent Human–Computer Interaction (HCI²): how far are we from attaining it?," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 1, 2008, pp. 168 - 187.

[24] M. Schröder et al., "The MARY TTS entry in the Blizzard Challenge 2008," *Proc. Blizzard Challenge*, Brisbane, AU: 2008.

[25] E. Bevacqua et al., "An expressive ECA showing complex emotions," *Proceedings of the AISB Annual Convention*, Newcastle, UK: 2007, pp. 208–216.

[26] D.C. Burnett, M.R. Walker, and A. Hunt, "Speech Synthesis Markup Language (SSML) Version 1.0," 2004; http://www.w3.org/TR/speech-synthesis/.

[27] I.S. Pandzic and R. Forchheimer, *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, J. Wiley & Sons, 2002.

[28] S. Kopp et al., "Towards a Common Framework for Multimodal Generation: The Behavior Markup Language," *Intelligent Virtual Agents*, 2006, pp. 205-217.

[29] D. Heylen et al., ed., *Proc. Workshop on Functional Markup Language at AAMAS'08*, Estoril, Portugal: 2008.

[30] M. Schröder et al., "What is most important for an Emotion Markup Language?," *Proc. Third Workshop Emotion and Computing, KI 2008*, Kaiserslautern, Germany: 2008.

[31] P. Baggia et al., "EMMA: Extensible MultiModal Annotation markup language," Dec. 2007; http://www.w3.org/TR/emma/.

[32] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese*1," *Journal of Pragmatics*, vol. 32, Jul. 2000, pp. 1177-1207.

[33] D. Bullock, "Adaptive neural models of queuing and timing in fluent action," *Trends in Cognitive Sciences*, 8, 2004, pp. 426-433.