SEMAINE

THE SENSITIVE AGENT PROJECT

D5a

SAL multimodal generation component with customised SAL characters and visual mimicking behaviour

2007 - 2013

Date: 22 December 2009

Dissemination level: Public

ICT project contract no.	211486
Project title	**SEMAINE** **Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression**
Contractual date of delivery	*31 December 2009*
Actual date of delivery	*22 December 2009*
Deliverable number	D5a
Deliverable title	SAL multimodal generation component with customised SAL characters and visual mimicking behaviour
Type	Demonstrator
Number of pages	22
WP contributing to the deliverable	WP 5
Responsible for task	Catherine Pelachaud (catherine.pelachaud@telecom-paristech.fr)
Author(s)	Elisabetta Bevacqua, Margaret McRorie, Satish Pammi, Catherine Pelachaud, Marc Schröder, Ian Sneddon, Etienne de Sevin.
EC Project Officer	Philippe Gelin

Table of Contents

1 Executive Summary

Sensitive Artificial Listeners (SAL) are virtual dialogue partners who, despite their very limited verbal understanding, intend to engage the user in a conversation by paying attention to the user's emotions and non-verbal expressions. The SAL characters have their own emotionally defined personality, and attempt to drag the user towards their dominant emotion, through a combination of verbal and non-verbal expression.

The SEMAINE system 2.0 is the first public demonstrator of the fully operational autonomous SAL system based on audiovisual analysis and synthesis. The present report is part of a group of reports describing various aspects of the SEMAINE system 2.0. The full list of reports is available from http://semaine.opendfki.de/wiki/SEMAINE-2.0.

This report describes the progress made in the multimodal generation components. It presents the state of the graphics engine that is responsible for the animation and visualization of the virtual agents. Four Sensitive Artificial Listeners have been created with specific facial models and expressive voices as well as their own lexicon and behavior characteristics.

2 System description

In this section we report on the development that was made on the output part of the Semaine architecture. We first report on changes made in the Greta's system. We then present the 4 SAL characters, their facial models and voices. We also report on some incremental developments that have been made in the systems. Finally we describe the vocal backchannels model followed by the representation of multimodal backchannels.

2.1 System development: Greta

2.1.1 General changes in Greta

We have fully integrated all the five Greta modules in the SEMAINE architecture: the listener intent planner, the listener action selection, the behaviour planner, the behaviour realise and the 3D player. This integration has implied many changes in the Greta architecture:

1. Greta's core

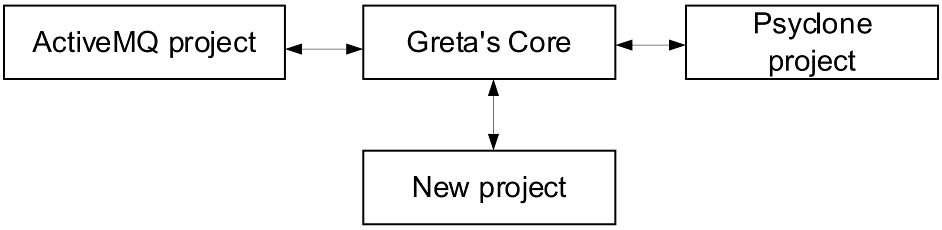

Figure 1. General schema of the Greta's architecture.

We separate the core of the Greta modules into core libraries from the communication API such as Psyclone or ActiveMQ (see figure 1). Greta system has been first developed using Psyclone communication API while in SEMAINE it is ActiveMQ that is used for communication API. In practice, we have two separate development projects (one for Psyclone and one for ActiveMQ) which call the same core libraries. All the common implementation is in the core libraries and the specific one relative to communication API is in the development projects.

2. Independence of TTS

In Greta architecture, we use external Text-To-Speech such as Festival, OpenMary or Acapela. The system is designed to be independent of the TTS and to be able to change it easily. The choice of the TTS, the language (be English, Italian, etc) and the voice is specified in an initialisation file.

3. XML namespaces

We introduce XML namespaces into FML and BML in the SEMAINE project to separate information about the speech part and the visual part. The latest is managed by Greta's architecture and the first one by the TTS (openmary). To be able to handle namespaces, we had to change our XML library as it did not support XML namespaces. We now use xerces. Thus openmary calculates automatically the information related to speech (through the Speech Preprocessor and Speech BML Realiser modules) and adds it in the FML and BML files with a specific XML namespace. Greta's architecture can read this information about them.

4. Logs and exceptions

In order to improve the debugging of the Greta core and of the development projects, we have added the possibility to have logs and exceptions. The latest ones allow us to find more easily the origin of program crashes and to exit them properly. With the logs, we can follow exactly what the program is doing during its execution. It is useful for doing software benchmarks or storing results.

2.1.2 Changes in Greta modules

In this section we describe the progress done within the SEMAINE Project on the Greta system. The whole system, presented in detail in the previous Deliverable, follows the SAIBA architecture (Vilhjalmsson et al, 2007); it is modular and distributed. Hereafter, we present each component, underlying the changes and the improvements.

1. Listener Intent Planner

The Listener Intent Planner (described in detail in D4a work-package) has been recently integrated in the SEMAINE architecture. It computes the agent's behaviour while being a listener conversing with a user. It decides *when* a backchannel signal should be emitted and selects *which* communicative intentions the agent should transmit through the signal. In order to generate the listener's signals, this module needs information about the user's behaviour: research has shown that there is a strong correlation between the triggering of a backchannel signal and the verbal and non verbal behaviours performed by the speaker (Maatman et al., 2005; Ward and Tsukahara, 2000). Then, to integrate the Listener Intent Planner in the SEMAINE architecture, we connected it with the input analysis applications.

The Listener Intent Planner is implemented in the ListenerIntentPlanner component in the SEMAINE framework. It receives information from the Topics semaine.data.state.agent, semaine.data.state.user.behaviour, semaine.data.state.dialog, semaine.data.state.context. Reactive/response backchannels are sent as FML file to the Topic semaine.data.action.candidate.function and mimicry as BML to the Topic semaine.data.action.candidate.behaviour.

2. Listener Action Selection

The Listener Action selection (described in detail in D4a work-package) (de Sevin and Pelachaud, 2009) receives all the candidate actions coming from the action proposers (Listener Intent Planner and Utterance Action Proposer). These candidate actions can be backchannels and utterances in FML or BML. The listener action selection received information about the turn-taking, the user interest level, the name of the character and the player call-backs from the SEMAINE architecture by subscribing to topics. The turn-taking allows differentiating when the agent is listening or speaking. A selection is made only in listener mode. The name of the character is used to adapt the behaviour selection depending on the agent's personality traits. The user interest level is estimated through visual and acoustic analysis and allows us to normalise the priorities of the candidate actions according to the behaviours of the user. Finally thanks to the player call-backs (see section 2.1.3), we can control the output of the system such as waiting that the current action is finished to be displayed to send a new one. If actions are received before the end of the current action, there are queued (include utterance actions) and used in the next choice. According to all these information, the listener action selection selects the most appropriate action to be displayed by the agent and send it in FML or BML to the output part of the SEMAINE architecture. Moreover the core of the listener action selection is implemented to be independent to the communication API such as ActiveMQ (see section 2.1.1).

The Listener Action Selection is implemented in the Action Selection component in the SEMAINE framework. It receives candidate FMLs from the Topic semaine.data.action.candidate.-function and BMLs from semaine.data.action.candidate.behaviour coming from Action Proposers. It also uses information from the Topics semaine.data.state.agent, semaine.data.state.user.behaviour, semaine.data.state.dialog, semaine.data.state.context and semaine.callback.output. Selected FMLs are sent to the Topic semaine.data.action.selected.function and selected BMLs to the Topic semaine.data.action.selected.behaviour.

3. Behaviour Planner

The Behaviour Planner takes as input both the agent's communicative intentions specified by the FML-APML language and some of the agent's characteristics. The main task of this component is to select, for each communicative intention to transmit, the adequate set of behaviours to display. All possible sets of behaviours for a given communicative intention are defined in a *lexicon*. To describe the agent's characteristics we adopted the concept of *baseline* proposed by Mancini in (Mancini and Pelachaud, 2008). The baseline is a set of numeric parameters in which two kinds of data are represented: the agent's modality preference and the agent's behaviour expressivity.

Within the SEMAINE project we define a lexicon and a baseline for each character according to its personality and emotional traits (see Section 2.2.2). Every time a character is called to interact with the user, both its lexicon and its baseline are automatically switched in the Behaviour Planner. In this way the generated animation varies according to the displayed agent. In order to avoid stalling the system while the proper data is charged, all baselines and lexicons are loaded when the system is launched.

The Behaviour Planner is implemented in the BehaviourPlanner component in the SEMAINE framework. It receives FMLs from the Topic semaine.data.action.selected.speechpreprocessed. It also uses information from the Topics semaine.data.state.agent and semaine.data.state.context. BMLs are sent to the Topic semaine.data.synthesis.plan.

4. Behaviour Realiser

The Behaviour Realiser is implemented in the BehaviourRealizer component in the SEMAINE framework. It receives BMLs from the Topics semaine.data.synthesis.plan.speechtimings and semaine.data.action.selected.behaviour. FAPs are sent to the Topic semaine.data.synthesis.lowlevel.video.FAP, BAPs to the Topic semaine.data.synthesis.lowlevel.video.BAP and commands to the Topic semaine.data.synthesis.lowlevel.video.command.

5. FAP-BAP Player

The FAP-BAP Player receives the animation and plays it in a graphic window. Facial and body configurations are described through respectively FAP and BAP frames. The new Player uses OGRE graphics engine and DirectX9 technology.

Within the SEMAINE project four different virtual agents can be displayed in the graphic window; the user can decide which agent she wants to interact with. For such a reason we implemented a dynamic loading of the four agents that allows the system to pass easily from one character to another when needed. Since each head is quite heavy (12300 triangles per mesh on average), the four agents are loaded in the memory when the system is launched. They are shown or hidden in the virtual word as needed. In this way the selected character can be displayed rapidly. In the previous version of the Player, the animation was automatically played by the virtual agent Greta, which was the only agent that could be visualized. Since four more agents have been added, the Player has been modified to send the animation to the agent that is actually displayed. Callbacks about FAPs, BAPs and audio are sent to the Topic semaine.callback.output.

The FAP-BAP Player is implemented in the PlayerOgre component in the SEMAINE framework. It receives FAPs from the Topic semaine.data.synthesis.lowlevel.video.FAP, BAPs from the Topic semaine.data.synthesis.lowlevel.video.BAP and audio files from the Topic semaine.data.synthesis.lowlevel.video.audio. It also uses information from the Topics semaine.data.state.context.

2.1.3 Incremental development work

Many implementation works have been done to enhance the functionality of the Greta modules in the SEMAINE architecture:

1. Player call-backs

We implement call-backs in the player in order to have information about what the player does and when. They correspond to the start and the end of audio, FAP and BAP. For example, the audio call-backs are used by the audio input to inactivate the microphone when the agent is speaking to avoid reverberations and to enhance the turn-taking detection. But the principal utility is to manage the output part of the agent. Indeed the player can display only one action at a time. So the current action has to be finished before the player can display the new one (we don't manage interruption yet). The call-backs allows the listener action selection to know when the selected action is started and finished to be displayed in order to manage the sending and the queuing of actions. After sending an action, the action selection waits until this action is finished to be displayed by the player before sending another one thanks to the call-backs. If actions are received during this time by the action selection, they are stacked in a queue and used in the next choice.

2. Message IDs

Some addition information has been added in the envelope of the SEMAINE message to know the history of its time creation and its propagation in the architecture. All the Greta modules have this type of information especially useful for the player call-backs. It is used for scheduling actions in order to verify if the current action displayed by the agent is the right one. It gives visibility in the architecture because we can know from which BML and FML, the FAP or the BAP comes from (for example, fml2_bml3_fap5) and how long takes the propagation. With these IDs, we can track easily the history of each file sent, thus enhancing the control in the SEMAINE architecture.

3. Release version

In order to run Greta's modules without installing all the programs and compiling them, a release version has been made. With these releases, Greta can normally run on every Windows computer and be easily distributed.

2.2 The four SAL characters

2.2.1 Conceptual specification of the characters

We are all familiar with the people in our daily lives demonstrating relatively stable, and often predictable, sets of behavioural characteristics. From such perceptions we automatically make judgements about the personalities of those we interact with. We need to be able to make the same kind of judgements about virtual agents. The credibility of such agents is dependent on them being perceived as coherent entities. To date, many of the behavioural characteristics that we use to make such automatic judgements (such as facial expressions, head and eye movements etc.) have been added to virtual agents in a way that is based, at best, on intuition. However, such intuitive construction of characters can be unconvincing and there tends to be little modelling of the most fundamental individual differences between humans – personality. We propose that the way in which personality dimensions affect various attributes of animated characters should reflect similar processes in the humans upon which they are modelled. If our virtual agents are to be capable of sustaining realistic interaction with human users, we need to consider how personality acts not only on the behavioural characteristics of these agents, but on their communicative styles. Our objective is to ensure that behavioural perceptions of a virtual agent credibly reflect the agent's 'actual' personality as prescribed. We thus use a solid theoretical basis to explain characterization of agents, and the conceptual specification of our characters is founded on sound psychological principles.

Trait models of personality assume that traits influence behaviour, and that they are fundamental properties of an individual. The five-factor model (McCrae et al., 1987) is a modern lexical approach, and posits five main personality dimensions – extraversion, neuroticism, openness to experience, agreeableness and conscientiousness. In comparison, Eysenck (Eysenck, 1976) developed a model based on traits which he believed were heritable and had a probable biological foundation. Likely personality traits were identified from clinical and experimental literature, and the three main traits which met these criteria were extraversion-introversion, neuroticism-emotional stability, and psychoticism. There is evidence of some form of theoretical integration between the two models. Eysenck's traits of extraversion and neuroticism are virtually identical to the similarly named dimensions of the 'Big Five', and psychoticism seems to correspond to agreeableness and conscientiousness combined – suggesting these traits may be components of psychoticism (Goldberg, 1993).

Eysenck attempted to provide causal explanations based on individual differences in nervous system functioning. His biological theory suggests that as extraverts are less cortically aroused than introverts, they should need more external stimulation and be more comfortable under arousing conditions. Highly neurotic individuals are predicted to show more autonomic nervous activity in stressful situations. Alternatively, M.W. Eysenck's Hypervigilance Theory argues that as 'highly anxious' people constantly look out for signs of threat, they will use many rapid eye movements, and attend selectively to threat-relevant stimuli (Eysenck and Byrne, 1992) . Psychoticism is less well understood, however Eysenck suggested psychoticism is linked to male hormones (e.g. testosterone) which influence impulsivity.

There is continued debate in the literature as to which of the two main personality models is more theoretically appropriate. What we have considered is which dimensions best reflect the various attributes of a virtual agent. The diversity created by the trait models provides a comprehensive framework, however modelling personality as a reflection of complex multivariate solutions might be difficult if virtual agents need more easily controlled parameters (Arya et al. 2006). Confining interpersonal behaviour to fewer dimensions allows for more effective management, and Eysenck's three-dimensional model arguably serves as an acceptable foundation. Its core dimensions of extraversion and neuroticism are undisputed and central to all major trait theories. Psychoticism is useful as it seems to reflect low agreeableness and low conscientiousness. Further benefits in adopting Eysenck's model are that its biological underpinnings can to some extent direct and justify specific response patterns of behaviour in developing characters of virtual agents.

These are the principles. In mapping the connections and considering how to translate these stable traits into personality-dependent actions, we focused on the links between impressions of personality and verbal/nonverbal behaviour, i.e. which behaviours actually affect a viewer's perception of personality. The five major categories typically used to classify nonverbal behaviour are facial expressions, eye and visual behaviour (e.g. gaze), kinesics, paralanguage, and proxemics.

Although the literature describing behaviours associated with particular human personalities is not couched in the same terms used to describe the agents, we have begun to use the human research to help us specify the design parameters for our characters. Four psychologically different affective/personality types have been created to elicit different types of emotion - each employing individual dialogue strategies, and displaying uniquely different responsive reactions. Poppy is outgoing (extraverted) and optimistic; Spike is angry and argumentative; Prudence is pragmatic and practical; and Obadiah is gloomy and depressed. Figure 2 portrays the 4 SAL facial models.

Poppy

The facial expressions of extraverts tend to be 'friendly', and the literature suggests that individuals are perceived as more sociable when smiling than with a neutral face (Borkenau and Liebler, 1992). We might further expect Poppy's facial appearance to be attractive. This relates to the 'what is beautiful is good' stereotype - positive personality attributions tend to be projected on to those possessing attractive faces (Feingold, 1992). Faces high in symmetry have similarly received significantly higher ratings of attractiveness, and facial symmetry is associated with personality attributes such as sociability, liveliness, and happiness (Fink et al., 2005). Drawing on Eysenck's theory of extraversion and arousal, Poppy would be characterized as having high levels of general activation. For example, extraverts tend to demonstrate more body movements, and display greater levels of facial activity (La France et al., 2004). Studies have also shown that extraversion is associated with greater levels of gesturing, more frequent head nods, and general speed of movement (Borkenau and Liebler, 1992). During conversation, extraverts tend to position themselves closer to others, and direct facial posture and eye contact is more likely to be maintained (Farabee et al. 1993). Regarding language and paralanguage, extraverts tend to talk more loudly and more repetitively, with fewer pauses, shorter silences, and less hesitations (Gill and Oberlander, 2002). Extraverts initiate more interactions, using more positive emotion words and informal style (McCroskey et al., 2001).

Spike

Spike's dispositional qualities of being angry and argumentative relate to Eysenck's third factor of psychoticism. This trait reflects hostility, and involves elements of aggression, coldness, impulsivity, and lack of empathy. Individuals high in psychoticism are more likely to be verbally aggressive, argumentative (thus low on agreeableness), and inappropriately assertive in communication (McCroskey et al., 2001). Eysenck's theoretical explanation for such behaviour proposed that psychoticism - like extraversion - reflected low cortical arousal, but was driven by abnormalities in neurotransmitter levels. He proposed links to levels of male hormones (e.g. testosterone), neurotransmitters and enzymes which influence impulsivity. According to Ekman and Friesen (Ekman and Friesen, 1975), facial expressions of anger are demonstrated with frowning eyebrows, staring eyes, and a closed mouth with depressed corners. Spike's facial appearance may thus be characterized by v-shaped eyebrows, with increased facial threat typified via prolonged direct eye gaze, wide eyes and open mouth (Tipples, 2007). Facial asymmetry and masculinity in male faces has been associated with lack of agreeableness and co-operation (Noor and Evans, 2003). When communicating, low scorers on agreeableness display less visual attention, but more visual dominance. Disagreeable individuals do less back-channelling, indicating they listen less to conversational partners (Smith, 1975).

Prudence

Prudence is characterized as practical and pragmatic. Her defining characteristic seems to be conscientiousness, however in expecting her movements to be 'measured' and 'economical', we might anticipate low impulsivity, and low psychoticism. In contrast to Spike, facial symmetry may thus be anticipated. Faces high in symmetry have received significantly higher ratings for competence and intelligence (Fink et al., 2005); and individuals wearing glasses tend to be rated as more intelligent (Argyle and McHenry, 1970). High eye contact has also been linked to competence, confidence, and self-esteem. Individuals who are thoughtful and reflective may show a predominance of upward looks (Exline and Winters, 1965). Those who are conscientiousness tend to avoid negations, negative emotion words and words reflecting discrepancies (e.g. should and would). Speech rate is positively correlated with perceived competence and conscientiousness.

Obadiah

Obadiah's defining features are gloominess and depression, which are characteristic of neuroticism – the enduring tendency to experience negative emotional states. This trait is related to pessimism, and in contrast to emotional stability, high neuroticism scorers tend to be frequently anxious, and are quicker to react to normal stress (Eysenck, 1976). As would be expected, full-face negative facial expression is directly related to neuroticism. However asymmetrically shaped faces also tend to be rated as more neurotic and less agreeable (Noor and Evans, 2003). Gaze avoidance and less eye contact are further cues for anxiety (Larsen and Schackelford, 1996). Neuroticism is associated with high levels of restlessness. This may be explained using Eysenck's prediction that neurotic individuals have a low activation threshold and show a greater degree of emotional arousal to minor stress. So, we might anticipate that Obadiah could get easily upset and fidget in social situations. As predicted by Hypervigilance Theory, Obadiah's character would also be expected to demonstrate more rapid eye movements (Eysenck and Byrne, 1992). This trait predicts a negative emotional tone. In contrast to Poppy, Obadiah would be expected to use more negative and fewer positive emotion words. High neuroticism scorers produce more self-references, and in conversation tend to have low, constant voice intensity (Pennebaker and King, 1999).

We have considered how behaviours associated with personality types in people may be adapted to develop characteristics of virtual agents. Four distinctive agents have been designed with a given personality each. For each of them, personality affects the agent's global behaviour quality as well as their backchannel productions (frequency and type of signals).

2.2.2 Personality lexicon and baseline

To model behavioural characteristics of virtual agents, we use the approach developed by (Mancini and Pelachaud, 2008) where an agent is defined by a baseline. The baseline describes how the agent behaves in terms of quality of movement and modalities used. It captures the global behaviour quality of the agent. The baseline specifies if an agent has the tendency to move slowly, with low amplitude or in a fast and hectic manner and if it displays mainly facial expression, gesture, head movement or torso. The first term refers to the execution of behaviours; it is also called behaviour expressivity or behaviour quality. The second term relates to the modalities the agent uses most to convey information. It is also called modality preferences. This parameter indicates if the agent uses mainly its face or its gesture to communicate. The baseline is defined as a set of numeric parameters: the agent's modality preference and the agent's behaviour expressivity. The modality preference refers to the agent's degree of preference in using each available modality (face, head, gaze, gesture and torso) to communicate while the behaviour expressivity is represented by a set of 6 parameters that influence the quality of the agent's movements as was proposed by (Hartmann et al., 2005): the frequency (OAC parameter), speed (TMP parameter), spatial volume (SPC parameter), energy (POW parameter), fluidity (FLD parameter), and repetitivity (REP parameter) of the non-verbal signals produced by the agent. These expressivity parameters are defined for each modality: one set of parameters for the head movements, another set for the facial expressions, and so on.

2.2.3 Introducing SAL characters in a virtual agent system

To introduce the SAL agents in our virtual agent system, we need to define a baseline for each of them. We determine the agents' modality preference and the expressivity on each modality according to the characters description proposed in the previous sections (see Table 1). Moreover, in our system we can model the agent's gaze behaviour by specifying the value of three gaze temporal parameters (see Table 2) that define the agent's tendency to look at the user, to look away from the user and to sustain mutual gaze (Pelachaud and Bilvi, 2003).

Poppy	**OAC**	**TMP**	**SPC**	**POW**	**FLD**	**REP**
face-head-gesture	high	medium	high	medium	medium	high
Spike	**OAC**	**TMP**	**SPC**	**POW**	**FLD**	**REP**
face-head-gesture	medium	medium	medium	high	medium	high
Prudence	**OAC**	**TMP**	**SPC**	**POW**	**FLD**	**REP**
face-head-gesture	medium	medium	medium	medium	medium	medium
Obadiah	**OAC**	**TMP**	**SPC**	**POW**	**FLD**	**REP**
face-head-gesture	medium	low	medium	medium	low	low

Table 1: General description of the baseline of each SAL character.

Gaze length	**Look at**	**Look away**	**Mutual gaze**
Poppy	medium	medium	high
Spike	high	low	high
Prudence	medium	medium	medium
Obadiah	low	high	low

Table 2: General description of the agent's tendency to look at the user.

We also need to define a lexicon for each agent. Such a lexicon describes how the agent conveys a given communicative intention through signals on the different modalities. We define a default baseline that is common to all virtual agents and a specific lexicon that is proper to each agent. This approach allows us to specify only the behaviours that vary from agent to agent. For example, the agents can use the same set of signals to transmit the communicative intention "ask", but a different set of signals to convey the intention “agree”. At present, we are determining which behaviours (and the signals they need to be composed of) differ from an agent to another.

2.2.4 Facial models

Four facial models have been created within the SEMAINE project to represent the 4 characters: Poppy, Prudence, Obadiah and Spike. The four agents have been designed taking into account their emotional traits and personality. QUB has also provided illustrations through drawing and pictures (from paintings and photos) of each SAL. From this illustrative data, we have discussed with a graphics designer to decide on the geometric models of the SAL. For example Spike’s facial appearance is characterized by v-shaped eyebrows, since he is the angry and argumentative one. The 4 new facial models can be seen in Figure 2. Head models have been created in format OBJ using Maya. In order to load the new heads in the Player we had to convert them in the right binary format used by Ogre. For such a purpose we used firstly the “Ogre meshes OgreXML exporter” script to export the OBJ files in OgreXML format (.mesh.xml) and then we applied the OgreXML-Converter to obtain the binary format.

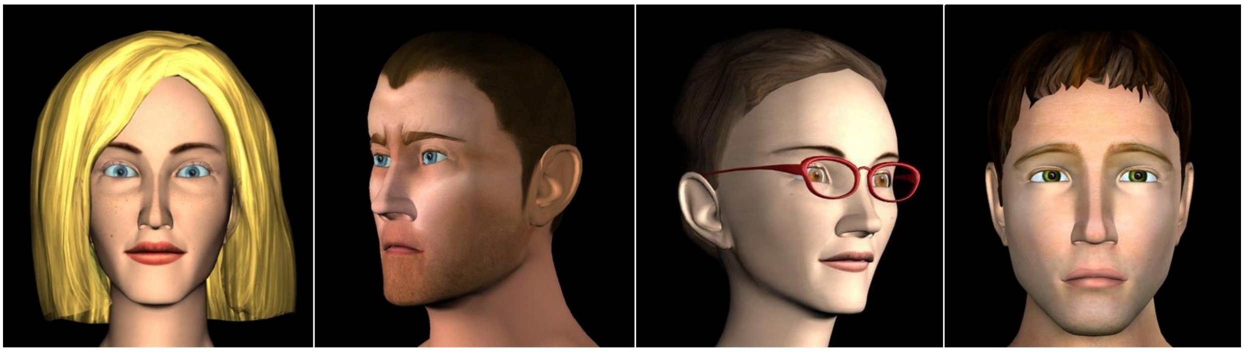

Figure 2. The 4 SAL agents: Poppy, Spike, Prudence and Obadiah.

The Player is MPEG-4 compliant (Ostermann, 2002). Before being able to animate the 4 facial models, we had to specify MPEG-4 animation parameters. In particular, we had to define the areas influenced by each FAP (facial animation parameter) and the position of the FDPs (facial definition parameters). For such a purpose we used our tool HeadTool. Figures 3 shows how the HeadTool has been used. Firstly the head model is loaded (Figure 3(a)), then the areas are defined by colouring all the triangles belonging to a given area (Figure 3(b)) and finally the FDPs are selected (the green point on the face in Figure 3(c)).

Since the HeadTool can load head models in VTX format, we had to create a converter from OgreXML to VTX ad hoc. Thanks to the definition of the areas and the FDPs the Player is able to display animations on the new agents. However, since each head has its own geometrical characteristics, for each FAP of each face we had to calibrate the amplitude and the intensity of the movement.

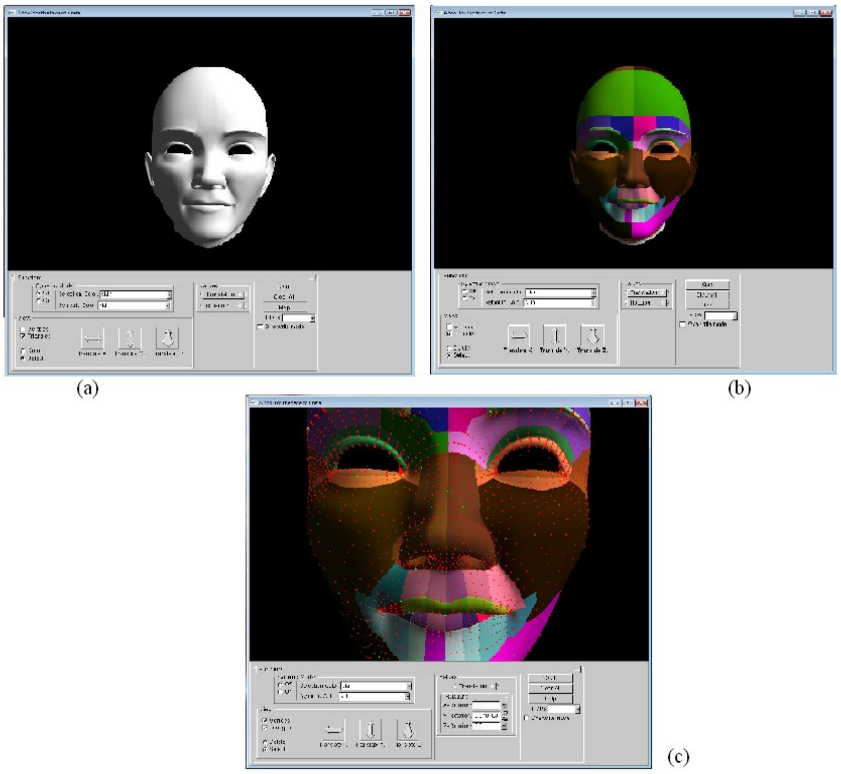

Figure 3. HeadTool examples. (a) the head is loaded in the application; (b) the areas are defined by colouring all the triangles that belong to a given area; (c) the FDPs are selected (the green points).

2.2.5 Expressive speech synthesis voices for the SAL characters

We created new British English speech synthesis voices for the four SAL characters. For each character, a professional actor was selected in a two-step casting process. The job was advertised on a web-based actor portal, and received 280 applications. Participants with a voice reel who seemed to have a suitable voice for one of the characters were invited to speak a small number of sentences from the respective recording script. Project members selected the voices that seemed to fit best with the facial models and the intended personality of the character.

Recordings were carried out in the Anechoic Chamber of University College London, throughout the course of one week. Each speaker produced around 150 sentences from the respective SAL speaker's script, and between 500 and 2000 sentences from the Wikipedia selected to optimise phonetic and prosodic coverage. The speakers were instructed to produce the Wikipedia sentences in the same speaking style as the domain-specific sentences. The combination of domain-specific and generic speech material yields “domain-oriented” (Schweitzer et al., 2003) voices: they sound very good within the target domain, but they can also produce arbitrary text with a reduced quality. The latter capability is important in the SEMAINE scenario since it is expected that new sentences will be needed as a result of the iterative evaluation of the system.

DFKI's voice creation toolkit (Schröder et al., 2009) was used to generate unit selection voices from the recorded material. The phonetic labels predicted by the text-to-speech system MARY were force-aligned and then sorted by a quality control algorithm (Pammi et al., 2009). A phonetician manually corrected the labels in the sentences that the quality control identified as potentially most problematic. Table 3 lists the amount of speech material and the proportion that was manually corrected for each voice. It can be seen that for Prudence, roughly twice the amount of speech material was recorded compared to the other voices. This was done because Prudence is the character whose voice is closest to an emotionally neutral voice, so it is the prime candidate for speaking unexpected text. A larger unit selection corpus is likely to increase the quality of such open-domain speech.

	Prudence	**Poppy**	**Spike**	**Obadiah**
sentences recorded	2224	783	1188	1014
sentences manually corrected	350	290	270	255

Table 3: Speech material used for creating the different synthetic voices

In addition to traditional speech synthesis material, we also recorded approximately 30 minutes of free dialogue with each actor in order to record listener vocalisations (Pammi & Schröder, 2009). Actor and interlocutor were located in different rooms so that their voices could be recorded onto separate audio channels; the actor was hearing the interlocutor through closed system headphones, to avoid leakage of the headphone output to the actor's microphone. Actors were instructed to participate in a free dialogue, but to take predominantly a listener role. We encouraged them to use “small sounds that are not words”, such as *mm-hm*, where it felt natural, in order to keep their interlocutor talking for as long as possible. However, they were also allowed to “say something” and therefore to become the speaker in the conversation where this “felt natural” to keep the dialogue going. One of three experimenters acted as the interlocutor.

The resulting speech synthesis voices are made publicly available as part of MARY TTS 4.0.0.

2.3 Vocal backchannels

MARY TTS was enhanced for the generation of non-verbal and quasi non-verbal vocalisations such as backchannels. A new element was introduced into the MARY-specific TTS markup format MaryXML. It allows a user to request a vocalisation based on the following criteria:

- meaning: the intended meaning of the vocalisation;
- intonation: the type of intonation contour used on the vocalisation;
- voice quality: the voice quality used with the vocalisation;
- name: a description of the segmental form of the vocalisation.

Table 4 lists the possible values currently supported for each of the criteria; note that not all values are available with every voice.

Attribute	Possible values
meaning	anger, sadness, amusement, happiness, contempt, certain, uncertain, agreeing, disagreeing, interested, uninterested, low-anticipation, high-anticipation, low-solidarity, high-solidarity, low-antagonism, high-antagonism
intonation	rising, falling, high, low
voicequality	modal, creaky, whispery, breathy, tense, lax
name	yeah, yes, mhmh, mhm, right, tsright, tsyeah, aha, (snort), (sigh), definitely, really, gosh, ah I see, oh god (gasp), yeah absolutely

Table 4: Values currently supported for each of the attributes of the `<vocalization>` *element in MaryXML. The lists of values are to be revised and extended in 2010.*

Figure 4 illustrates the syntax of the new markup.

```
<?xml version="1.0" encoding="UTF-8" ?>
<maryxml version="0.5" xmlns="http://mary.dfki.de/2002/MaryXML" xml:lang="en-
GB">
  <voice name="dfki-prudence">
    <vocalization name="right" meaning="accept"/>
    <s>My name is Prudence.</s>
  </voice>
</maryxml>
```

Figure 4: Example of MaryXML markup requesting the generation of a vocalization followed by a verbal utterance.

All of the attributes of the vocalization tag are optional; if an attribute is not given, this means that the search is not constrained on that level. MARY TTS will look up available vocalisations for the given speaker and will generate the most appropriate vocalisation found for the request. As for normal text to speech, both audio and realised durations of allophones can be generated. For vocalisations that do not have a clear allophonic structure, such as sighs or laughs, the most similar allophones are returned in timing information formats such as REALISED_ACOUSTPARAMS or REALISED_DURATIONS.

As of version 4.0.0, the functionality of vocalisation lookup is implemented and working, but the annotation of the meaning of vocalisations for the four characters is still ongoing. For this reason, only a limited set of vocalisations can currently be requested.

2.4 Multimodal backchannels

As explained in Section 2.1.2, the Listener Intent Planner can generate three types of backchannels: mimicry, reactive and response. All these types of signals can be displayed by the agent through different modalities, like face, head, gaze, voice and so on. Mimicry backchannels can be written directly in BML language, since they specify the non-verbal signals, performed by the user that the agent has to mimic. Response and reactive backchannels are written in FML language, describing the communicative functions that the agent aims to convey through a backchannel signal. For example, the response backchannel that wants to transmit the communicative function “agreement” is generated by the Listener Intent Planner in the following FML (see figure 5):

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
```

```
<fml-apml>
 <fml xmlns="http://www.mindmakers.org/fml" id=”fml1”>
      <backchannel id="b1" type="agreement" start="0.0" end="1.5" importance="1.0"/>
</fml>
</fml-apml>
```

Figure 5: Example of FML generated by the Listener Intent Planner.

From the FML tag, the Behaviour Planner generates automatically a multimodal set of behavioural signals written in BML (see section 2.2.2), taking into account the agent's baseline and its lexicon. For example, for Poppy, the happy and optimistic one, the Behavior Planner could generate the following BML (see figure 6):

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<bml xmlns="http://www.mindmakers.org/projects/BML">

<head id="backchannel-0" start="0.00" end="1.50" direction="RIGHT" type="NOD">
  <description level="1" type="gretabml">
    <reference>head=head_nod</reference>
      <FLD.value>0.00</FLD.value>
      <OAC.value>0.80</OAC.value>
      <PWR.value>0.20</PWR.value>
      <REP.value>0.00</REP.value>
      <SPC.value>0.90</SPC.value>
      <TMP.value>0.80</TMP.value>
      <preference.value>0.90</preference.value>
  </description>
</head>

<face id="backchannel-1" start="0.00" end="1.50" shape="flat" type="MOUTH">
  <description level="1" type="gretabml">
    <reference>eyes=smile</reference>
      <FLD.value>0.50</FLD.value>
      <OAC.value>0.80</OAC.value>
      <PWR.value>0.20</PWR.value>
      <REP.value>0.00</REP.value>
      <SPC.value>0.90</SPC.value>
      <TMP.value>0.80</TMP.value>
      <preference.value>0.70</preference.value>
 </description>
</face>

<speech id="backchannel-2" language="en-GB" text="yes" type="application/wav"
voice="dfki-poppy">
  <vocalization xmlns="http://mary.dfki.de/2002/MaryXML/" intonation="rising"
meaning="friendly" name="yeah" voicequality="breathy"/>
    <description level="1" type="gretabml"/>
      yeah
</speech>
</bml>
```

Figure 6: Example of BML generated by the Behavior Planner for Poppy.

Poppy smiles and nods while saying "yeah" in a friendly way. For each behavioural signal on the non-verbal modalities, the expressivity parameters are specified. These parameters define the quality of the agent's movement. In our example, Poppy will perform a wide and fast head nod (SPC=0.90 and TMP=0.80, where SPC means spatial and TMP temporal).

Starting from the same FML, the Behaviour Planner could generate a different set of behavioural signals for Spike who, being aggressive and angry, has a different baseline and a different lexicon. For example the resulting BML could be (see figure 7):

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<bml xmlns="http://www.mindmakers.org/projects/BML">

<head id="backchannel-0" start="0.00" end="1.50" direction="RIGHT" type="NOD">
  <description level="1" type="gretabml">
    <reference>head=head_nod</reference>
      <FLD.value>0.00</FLD.value>
      <OAC.value>0.80</OAC.value>
      <PWR.value>0.80</PWR.value>
      <REP.value>0.00</REP.value>
      <SPC.value>0.20</SPC.value>
      <TMP.value>0.70</TMP.value>
      <preference.value>0.90</preference.value>
  </description>
</head>

<face id="backchannel-1" start="0.00" end="1.50" shape="flat" type="EYES">
  <description level="1" type="gretabml">
    <reference>eyes=raise_eyebrows</reference>
      <FLD.value>0.50</FLD.value>
      <OAC.value>0.80</OAC.value>
      <PWR.value>0.80</PWR.value>
      <REP.value>0.00</REP.value>
      <SPC.value>0.30</SPC.value>
      <TMP.value>0.70</TMP.value>
      <preference.value>0.70</preference.value>
 </description>
</face>

<speech id="backchannel-2" language="en-GB" text="m-mh" type="application/wav"
voice="dfki-spike">
  <vocalization xmlns="http://mary.dfki.de/2002/MaryXML/" intonation="rising"
meaning="unfriendly" name="m-mh" voicequality="breathy"/>
    <description level="1" type="gretabml"/>
      m-mh
</speech>
</bml>
```

Figure 7: Example of BML generated by the Behavior Planner for Spike.

Spike raises his eyebrows and nods while saying “m-mh” in an unfriendly way. Differently from Poppy, Spike performs narrow and powerful movements.

3 License and availability

MARY TTS 4.0.0 is available from http://mary.dfki.de. The TTS system is licensed under the Lesser GNU General Public License, LGPL, http://www.gnu.org/licenses/lgpl-3.0-standalone.html.

The speech synthesis voices dfki-prudence, dfki-poppy, dfki-spike and dfki-obadiah can be installed using the MARY component installer which is part of the MARY TTS installation process. The voices are distributed under the terms of the Creative Commons Attribution-NoDerivatives license, http://mary.dfki.de/download/by-nd-3.0.html.

MARY TTS and the four synthesis voices are part of the SEMAINE 2.0 system release.

Greta is available from http://www.tsi.enst.fr/~pelachau/Greta/. It is licensed under GPL licence. Greta and the four facial models are part of the SEMAINE 2.0 system release.

References

(Argyle and McHenry, 1970) Argyle, M., McHenry, R.: Do spectacles really affect judgements of intelligence? British Journal of Social and Clinical Psychology, 10(1), 27-29 (1970)

(Arya et al. 2006) Arya, A., Jefferies, L., Enns, J.T., DiPaola: Facial actions as visual cues for personality. Computer Animation and Virtual Worlds, 17, 1-12. 2006.

(Bevacqua et al., 2008) Bevacqua, E., Mancini, M., and Pelachaud, C. A listening agent exhibiting variable behaviour. In Prendinger, H., Lester, J. C., and Ishizuka, M., editors, Proceedings of 8th International Conference on Intelligent Virtual Agents, volume 5208 of Lecture Notes in Computer Science, pages 262-269, Tokyo, Japan. Springer, 2008.

(Borkenau and Liebler, 1992) Borkenau, P., Liebler, A.: Trait inferences: Sources of validity at zero acquaintance. Journal of Personality and Social Psychology, 62, 645-657 (1992)

(de Sevin and Pelachaud, 2009) de Sevin, E. and Pelachaud, C.: "Real-time Backchannel Selection for ECAs according to User's Level of Interest". In Proceedings of Intelligent Virtual Agents 2009, IVA'09, Amsterdam, Holland. 2009.

(Exline and Winters, 1965) Exline, R.V., Winters, L.C.: Affect relations and mutual gaze in dyads. In S. Tomkins, and C. Izzard (Eds.), Affect, Cognition and Personality. Springer, New York (1965)

(Eysenck, 1976) Eysenck, H.J.: The Measurement of Personality. Lancaster: Medical and Technical Publishers. 1976.

(Eysenck and Byrne, 1992) Eysenck, M.W., Byrne, A.: Anxiety and susceptibility to distraction. Personality and Individual Differences, 13, 793-798 (1992)

(Ekman and Friesen, 1975) Ekman, P and Friesen, W. Unmasking the Face. A Guide to Recognizing Emotions from Facial Clues. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

(Farabee et al. 1993) Farabee, D., Nelson, R., Spence, R.: Psychosocial profiles of criminal justice- and non-criminal justice-referred clients in treatment. Criminal Justice and Behaviour, 20, 336-346 (1993)

(Feingold, 1992) Feingold, A.: Good-looking people are not what we think. Psychological Bulletin, 111, 304-341 (1992)

(Fink et al., 2005) Fink, N., Neave, N., Manning J.T., Grammar, K.: Facial symmetry and the big-five personality factors. Personality and Individual Differences, 39, 523-529 (2005)

(Gill and Oberlander, 2002) Gill, A., Oberlander, J.: Taking care of the linguistic features of Extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 363-368 (2002)

(Goldberg, 1993) Goldberg, L.R.: The structure of phenotypic personality traits. American Psychologist, 84, 26-34. 1993.

(Hartmann et al., 2005) Hartmann, B, Mancini, M, Buisine, S, and Pelachaud, C. Design and evaluation of expressive gesture synthesis for embodied conversational agents. In 3th International Joint Conference on Autonomous Agents & Multi-Agent Systems, Utretch, 2005.

(La France et al., 2004) La France, B., Heisel, A., Beatty, M. : Is there empirical evidence for a nonverbal profile of extraversion ? A meta-analysis and critique of the literature. Communication Monographs, 71, 28-48 (2004)

(Larsen and Schackelford, 1996) Larsen, R.J., Schackelford, T.: Gaze avoidance: personality and social judgements of people who avoid direct face-to-face contact. Personality and Individual Differences, 21, 907-917 (1996)

(McCrae et al., 1987) McCrae, R.R., Costa, P.T.: Validation of the five-factor model of personality across instruments and observers. Journal of Personality and Social Psychology, 52, 81-90. 1987.

(McCroskey et al., 2001) McCroskey, J, Heisel, A, and Richmond, V. Eysenck's big three and communication traits: three correlational studies. Communication Monographs, 68:360–386, 2001.

(Maatman et al., 2005) Maatman, R, Gratch, J, and Marsella, S. Natural behavior of a listening agent. In 5th International Conference on Interactive Virtual Agents. Kos, Greece, 2005.

(Mancini and Pelachaud, 2007) Mancini, M and Pelachaud, C. Dynamic behavior qualifiers for conversational agents. In Pelachaud, C, Martin, J.-C, André, E, Chollet, G, Karpouzis, K, and Pel´e, D, editors, Proceedings of 7th International Conference on Intelligent Virtual Agents, volume 4722 of Lecture Notes in Computer Science, pages 112–124, Paris, France, 2007. Springer.

(Mancini and Pelachaud, 2008) Mancini, M and Pelachaud, C. Distinctiveness in multimodal behaviors. In Padgham, L, Parkes, D, Müller, J, and Parsons, S, editors, Proceedings of Conference on Autonomous Agents and Multi-Agent Systems (AAMAS08), 2008.

(Noor and Evans, 2003) Noor, F and Evans, D. The effect of facial symmetry on perceptions of personality and attractiveness. Journal of Research in Personality, 37:339-347, 2003.

(Ostermann, 2002) Ostermann, J. Face animation in MPEG-4. In Pandzic, I. And Forchheimer, R., editors, MPEG-4 Facial Animation - The Standard Implementation and Applications, pages 17-55. Wiley, England, 2002.

(Pammi & Schröder, 2009) Pammi, S., & Schröder, M. (2009). Annotating meaning of listener vocalizations for speech synthesis. In Proc. International Conference on Affective Computing & Intelligent Interaction. Amsterdam, The Netherlands: IEEE.

(Pammi et al., 2009) Pammi, S., Charfuelan, M., & Schröder, M. (2009). Quality control of automatic labelling using HMM-based synthesis. In Proc ICASSP 2009. Taipei, Taiwan.

(Pelachaud and Bilvi, 2003) Pelachaud, C and Bilvi, M. Modelling gaze behavior for conversational agents. In IVA03 International Working Conference on Intelligent Virtual Agents, volume LNAI 2792, pages 15–17, Germany, 2003. Springer, 2003.

(Pennebaker and King, 1999) Pennebaker, J., King, L. Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77, 1296-1312 (1999)

(Smith, 1975) Smith, B, Brown, B, Strong, W, and Rencher, A. Effects of speech rate on personality perceptions. Language and Speech, 18:145-152, 1975.

(Schröder et al., 2009) Schröder, M., Pammi, S., & Türk, O. (2009). Multilingual MARY TTS participation in the Blizzard Challenge 2009. In Blizzard Challenge 2009. Edinburgh, UK.

(Schweitzer et al., 2003) Schweitzer, A., Braunschweiler, N., Klankert, T., Möbius, B., & Säuberlich, B. (2003). Restricted unlimited domain synthesis. In Proc. Eurospeech 2003. Geneva, Switzerland.

(Tipples, 2007) Tipples, J. Wide eyes and an open mouth enhance facial threat. Cognition and Emotion, 21:535 557, 2007.

(Vilhjálmsson et al., 2007) Vilhjálmsson, H. H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., Ruttkay, Z., Thórisson, K. R., van Welbergen, H., and van der Werf, R. J. The Behavior Markup Language: Recent developments and challenges. In Pelachaud, C., Martin, J.-C., Andr, E., Chollet, G., Karpouzis, K., and Pelé, D., editors, Proceedings of 7th International Conference on Intelligent Virtual Agents, volume 4722 of Lecture Notes in Computer Science, pages 99-111, Paris, France. Springer, 2007.

(Ward and Tsukahara, 2000) Ward, N and Tsukahara, W. Prosodic features which cue back-channel responses in English and japanese. Journal of Pragmatics, 23:1177–1207, 2000.