SEMAINE

THE SENSITIVE AGENT PROJECT

D5b

SAL multimodal generation component optimised for real-time behaviour

2007 - 2013

Date: 24 September 2010

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 August 2010*
Actual date of delivery	*24 September 2010*
Deliverable number	D5b
Deliverable title	SAL multimodal generation component optimised for real-time behaviour
Type	Demonstrator
Number of pages	15
WP contributing to the deliverable	WP 5
Responsible for task	Catherine Pelachaud (catherine.pelachaud@telecom-paristech.fr)
Author(s)	Elisabetta Bevacqua, Satish Pammi, Catherine Pelachaud, Marc Schröder, 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.

This report is part of the series of reports describing the implementation of SAL in system SEMAINE-3.0. The software described, and the full set of reports, can be downloaded from http://semaine.opendfki.de/wiki/SEMAINE-3.0/.

This report describes the progress done in the various modules of the multimodal agent architecture. We also present evaluation studies we conducted.

2 Functionality of the components

This section describes the functionality of the components in the SAL system. The possibilities to configure and reuse the components as parts of a research toolbox will be published as deliverable D7e in December 2010.

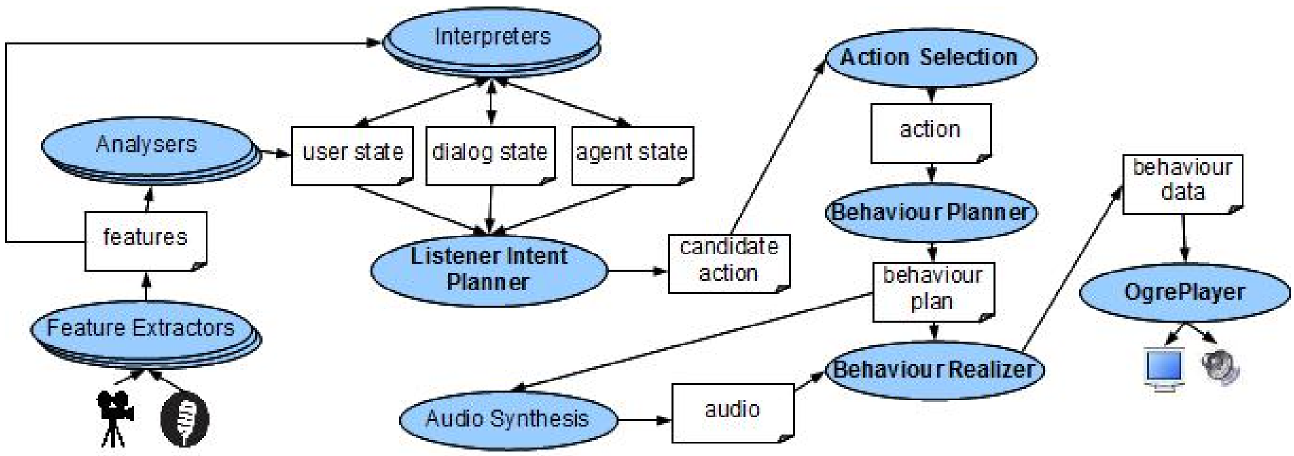

Figure 1. System architecture.

Figure 1 shows the architecture of the SEMAINE system that generates the agent's behaviour and the final visual output (the modules described here are written in bold). The whole architecture follows the SAIBA standard (Vilhjalmsson et al, 2007); it is modular and distributed. Hereafter, we present each component, underlying the changes and the improvements done in the last year.

2.1. Listener Intent Planner

The Listener Intent Planner module is in charge of the computation of the agent's behaviours while being a listener when conversing with a user. This component is able to generate two types of backchannel: response and mimicry. Response backchannels are the expression of what the listener thinks about the speaker's speech, we use Allwood's and Poggi's taxonomies of communicative functions of backchannels (Allwood et al., 1993, Poggi, 2007): understanding and attitudinal reactions (liking, accepting, agreeing, believing, being interested). Mimicry backchannels derive from the imitation of some behaviours performed by the speaker, like facial expressions, head movements and so on. Mimicry of behaviours may happen when the interactants are fully engaged in an interaction (Lakin et al., 2003). The Listener Intent Planner decide also when a backchannel should be triggered according to the speaker's acoustic and visual behaviour (Maatman et al., 2005,Ward and Tsukahara, 2000). The module has been presented in detail in the previous Deliverables.

The Listener Intent Planner was integrated in the SEMAINE architecture by connecting 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. Mimicry/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.

Progress in the Listener Intent Planner. To determine which communicative functions the agent will transmit through a response backchannel, the system needs to know what the agent *thinks* about the user's speech, for example if it agrees, refuses, likes the content of the message. This information, stored in the agent's mental state, is now computed by the dialogue manager and sent on the Topic semaine.data.agent.state. Recently we upgraded the Listener Intent Planner to receive and store the modifications of the agent's mental state. Moreover new communicative functions have been added in the agent's mental state to introduce the backchannel vocalizations studied in WP3 (see Deliverable X). Since the listener functions are not explicitly defined in the source code but specified in an XML-based language, the set of communicative functions has been easily extended by adding the new functions in the XML file. Then, it is the module that automatically creates the necessary internal structures to manage each function.

2.2. Listener Action Selection

The Action Selection (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 action selection received information about the turntaking, the user interest level (that is is a good indicator of the success of the interaction (Peters et al., 2005)), the name of the character and the player call-backs from the SEMAINE architecture by subscribing to topics. All these informations are used to computed the selection. The module has been presented in detail in previous Deliverables.

The action selection has been 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.

Progress in the Listener Action Selection. Recently this module has been modified to take into account the agent's personality in the selection process. In SEMAINE project, four SAL agents are designed with their own personality traits. Poppy is outgoing and cheerful; Spike is aggressive and argumentative; Prudence is reliable and pragmatic; and Obadiah is pessimistic and gloomy (see figure 2). We can place the four SAL agents in Eysenck's two dimensional representation (see figure 2) which are extroversion and neuroticism (emotional stability) (Eysenck, 1991).

The four SAL agents are placed in the dimensional space defined by Eysenck’s representation of personality. Values for mimicry tendency and BC frequency are defined according to their position in this space. For example, Obadiah which is pessimistic, performs few backchannels (-0.75) and sometimes some mimicry (-0.6). We obtain these values for the four personalities (see table 1). These agent's characteristics have an effect both on the type and the frequency of backchannel signals.

Backchannel types. Backchannel selection is event-based and is done in real-time. The algorithm follows a 2 steps process: it first determines when to do a backchannel (trigger phase) and then it selects which signal to display (selection phase). Only one action can be displayed by the ECA at a time and the Listener Action Selection receives continuously candidate backchannels. When the

ECA is already displaying a backchannel, no choices are made. The action selection algorithm waits until the display of the current backchannel is over before selecting another one to be displayed. These candidate backchannels received during this time are queued and used during the next selection pass. The listener Action Selection receives potential backchannel signals to display from the trigger module. These signals are characterized by a priority level. This value depends on the personality description of the agent; more particularly on its degree of neuroticism (Eysenck and Eysenck, 1978). A highly emotionally stable agent shows more mimicry behaviours (Chartrand et al., 2005) while a highly emotionally unstable agent shows more responsive behaviors (McCroskey et al., 2001, Noor and Evans, 2003). The priority associated to each backchannel action is computed according to its type (mimicry or responsive) and to the agent's personality (degree of neuroticism).

Backchannel frequency. Based on a theoretical model (Eysenck, 1991), we establish a correlation between the extroversion dimension and the frequency of backchannels (Borkenau and Liebler, 1992) From the videos analysis of SEMAINE corpus collected in QUB, we computed the backchannel frequency: Poppy's BC frequency is 20% higher than Spike's, Spike's BC frequency is 50% higher than Prudence's and Prudence's BC frequency is 30% higher than Obadiah's. The value of the frequency is deduced from our model (see section 3.2). For example, the value for Poppy (extravert) is 0.95 which means that the large majority of backchannels will be displayed (La France et al., 2004). One the other end, the value for Obadiah (introvert) is -0.75 which means only 25% of the backchannels will be displayed (Smith et al., 1975). When the Listener Action Selection receives a potential backchannel (mimicry or response backchannel), it calculates a probability in order to determine if the backchannel will be displayed or not, based on the degree of the agent's extroversion. If not, the backchannel is not queued by the Listener Action Selection.

The Action Selection algorithm has been evaluated in a perceptive study reported in the Evaluation Section.

•

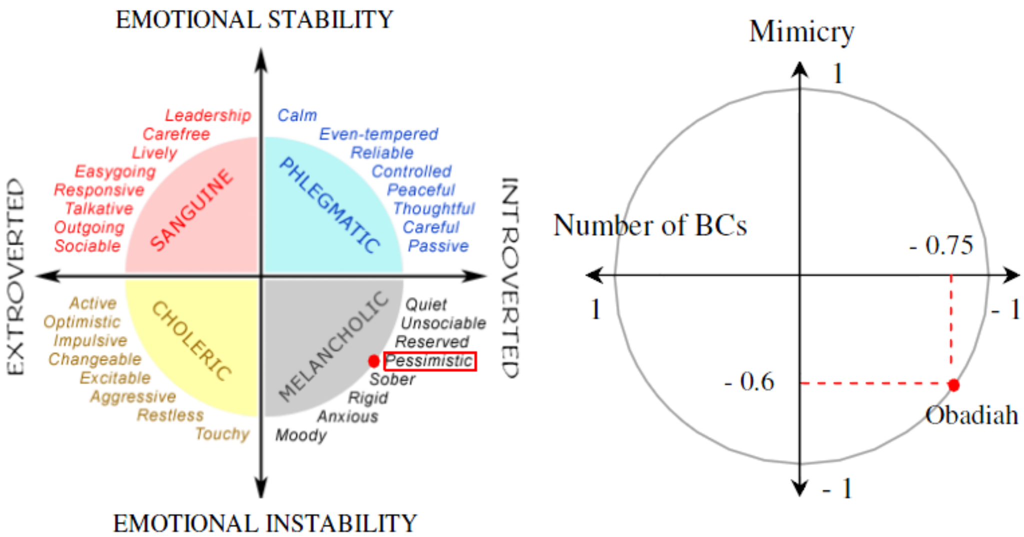

Figure 2. Eysenck's two dimensional representation and our hypothesis of its implication on mimicry and number of backchannels. Example of deduction for Obadiah.

	Obadiah (pessimistic)	Poppy (outgoing)	Prudence (reliable)	Spike (aggressive)
BC type (mimicry)	- 0.6	0.25	0.90	- 0.85
BC frequency	- 0.85	0.95	- 0.5	0.55

Table 1. Setting of BC type and frequency for the four SAL agents.

2.3. 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 according to the agent's characteristics described in its baseline, that is its modality preference and its behaviour expressivity (Mancini and Pelachaud, 2008). The module has been presented in detail in the previous Deliverables.

The Behaviour Planner has been 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.

Progress in the Behaviour Planner. All possible sets of behaviours for a given communicative intention are defined in a lexicon. Recently we defined a lexicon for each character to take into account their different personalities and then their different way of communicating. The four lexicons have been created analysing the videos collected in the SEMAINE database. The new lexicons allows the system to generate a more various and appropriate behaviour for the agent according to its personality traits.

We also introduced new communicative functions and consequently we extended the lexicons to contain the behaviour set for each new function. Particularly, new functions have been introduced for the agent while in the role of the listener, like antagonism, solidarity (see XXX).

Backchannel vocalizations have also been added in the lexicon. Through recent works we evaluate multimodal backchannel signals (that is signals composed by visual and vocal cues) to understand how users interpret them and then how we can use them to enrich the agent's behaviour while listening. To introduce these vocal signals in our system we added a new modality in the lexicon, the “speech” modality. Such a modality specifies for each listener communicative function which vocalizations can be emitted together with other visual signals to transmit a given function. Figure 3 shows an example of the behaviour set for the listener communicative function “agreement” for the agent Prudence. The fourth signal belongs to the speech modality and it specifies all the vocalizations the agent can emitted by Prudence to transmit agreement. The signals are interchangeable, with a certain probability, as specified in the alternative tags.

```
<behaviorset name="backchannel-agreement">
 <signals>
  <signal id="1" name="head=head_nod" modality="head">
   <alternative name="head=head_nod1" probability="0.3"/>
  </signal>
  <signal id="2" name="faceexp=prudence-agreement" modality="face">
   <alternative name="faceexp=close-eyes" probability="0.2"/>
   <alternative name="mouth=little_smile_open" probability="0.2"/>
  </signal>
  <signal id="3" name="gaze=look_at" modality="gaze"/>
  <signal id="4" name="text" modality="speech" content="yes" intonation="rising" voicequality="tense"
                   meaning="agreeing">
        <alternative name="text" content="yes" intonation="rising" voicequality="modal" meaning="agreeing"
                   probability="0.2"/>
        <alternative name="text" content="tsyeah" intonation="rising" voicequality="modal" meaning="agreeing"
                   probability="0.1"/>
        <alternative name="text" content="tsright" intonation="rising" voicequality="modal" meaning="agreeing"
                   probability="0.1"/>
        <alternative name="text" content="right" intonation="rising" voicequality="modal" meaning="agreeing"
                   probability="0.1"/>
        <alternative name="text" content="alright" intonation="rising" voicequality="modal" meaning="agreeing"
                   probability="0.1"/>
        <alternative name="text" content="yeah" intonation="rising" voicequality="breathy" meaning="agreeing"
                   probability="0.1"/>
        <alternative name="text" content="yeah_that's_true" intonation="rising" voicequality="modal"
                   meaning="agreeing" probability="0.1"/>
  </signal>

 </signals>
 <constraints>
  <core>
   <item id="1"/>
  </core>
  <rules>
   <implication>
    <ifpresent id="2"/>
    <thenpresent id="3"/>
   </implication>
  </rules>
 </constraints>
</behaviorset>
```

Figure 3. Example of behaviour set of the listener's communicative function for Prudence.

2.4. 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.

2.5. Secondary path: Prepare and trigger branch

A fundamental property of a real-time system is the speed of reaction. Due to the heavy amount of computation and physic limitation our system is not always able to generate a response in an acceptable interval of time. In order to speed up the system we duplicate both the Behaviour Planner and the Behaviour Realizer. Together they create a secondary path that works in parallel with the first one to generate suitable actions. These actions are stored in a queue and played when the agent is due to play an action but the first path has not yet finished to generate the requested one.

The details of this mechanism are described in D1d, sections 2.3 and 2.4.

2.6. 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 Player is based on OGRE graphics engine that can use either DirectX9 technology or OpenGL libraries. 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. These agents are loaded dynamically, 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. The Player sends directly 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 OgrePlayer 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.

Progress in the FAP-BAP Player. To improve the Player, two queues have been implemented. The first queue, called “waiting queue”, is used to store data about animations that have been computed by the previous modules of the system. These animations are moved in the second queue, called “ready queue”, when the minimum set of needed information have been received (like FAPs, BAPs, audio and start time). The animations in the ready queue are played as their start time corresponds to the clock of the system. The animations in the waiting queue are ordered according to a priority value and they do not remain indefinitely in the queue. A life time is associated to each action which is discarded as soon as their life time is exceeded.

3 Quality assessment

This section describes an assessment of the quality on the technical and component level. An assessment of the psychological quality of interactions with the overall system will be published separately, as deliverable report D6d, in December 2010.

3.1 Perceptual evaluation

To asses the quality of our components we performed perceptive evaluations. As for the Action Selection module we tested how backchannels influence the perception of the agent's personality. We evaluated our model with two hypotheses: backchannel frequency is linked with the extroversion dimension and backchannel type with the neuroticism dimension. The results show that the first hypothesis is partially verified for outgoing and reliable personalities. The second hypothesis is verified only for the pessimistic personality. For this hypothesis, the terms used for personality need to be clarified in order to the participants understand them correctly. Although more evaluation tests are needed, the selection of backchannel type and frequency by the Listener Action Selection according to our model help to express some personalities displayed by the ECA. Further information on this evaluation study and its results can be found in (de Sevin et al., 2010).

Another evaluation has been conducted in order to build the lexicons. Such a study aims at defining a set of multimodal backchannel signals and tests how they are interpreted by subjects when displayed context-free, that is without knowing the discursive context of the speaker's speech. For multimodal signals we mean signals that contain both visual and vocal cues. Through the evaluation we saw also that the meaning conveyed by a multimodal backchannel cannot be simply inferred by the meaning of each visual and vocal cues that compose it and that certain signals can be very dependant to the content of the speaker's speech, like signals composed by visual and vocal cues that have strongly opposite meanings. In (Bevacqua et al., 2010) details on this perceptual study are provided.

3.2 Generation performance

The speed improvement due to the secondary prepare-and-trigger branch for generating agent behaviour can be measured in the live system. Table 1 shows measures from a number of test runs of the full audio-visual live system in different test configurations.

Test setup	**Direct branch**	**Prepare-and-trigger branch**
Distributed system, no cache in MARY TTS	687 ms (n=82)	7 ms (n=16)
Distributed system, cache enabled in MARY TTS	201 ms (n=80)	5 ms (n=22)
Single PC, no cache in MARY TTS	846 ms (n=46)	7 ms (n=22)
Single PC, cache enabled in MARY TTS	407 ms (n=41)	18 ms (n=39)

Table 1: Median time-to-animation in various system setups, in milliseconds.

A number of things can be seen from Table 1. First, depending on the course of the dialogue, the number of utterances that are generated using the direct branch is several times larger than the number of utterances that are generated using the prepare-and-trigger branch. This is to some extent expected behaviour, because the user's behaviour may or may not be in line with the system's predictions, so that any prepared utterances may or may not be appropriate.

Second, the TTS cache in the MARY TTS system helps to reduce the time in the direct prepare branch to less than half. The remaining time is mainly the preparation of the visual animation.

Third, distributing the load across two machines (one running the dialogue, speech synthesis and speech input components, the other running video input and output) also helps speed up the generation of output. The main reason for this seems to be that all processes are rather resource intensive, so that the single PC is under heavy load.

Finally, and most importantly for this investigation, when the prepare-and-trigger branch can be used for generating a system output, this dramatically improves the time from the decision to generate a behaviour to its start. Latencies drop from roughly half a second to roughly 10 ms. This is indeed much more responsive.

In conclusion, the prepare-and-trigger architecture achieves very well its aim of improving the reaction speed of the system for those utterances that can be prepared ahead of time. However, this is still less than half the utterances. It is worth investigating whether the number of utterances that can be prepared ahead of time can be increased further.

4 License and availability

MARY TTS 4.1.1 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 3.0 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 3.0 system release.

References

(Allwood et al., 1993) Allwood, J., Nivre, J., and E., A. (1993). On the semantics and pragmatics of linguistic feedback. *Semantics*, 9(1).

(Bevacqua et al., 2007) E. Bevacqua, D. Heylen, M. Tellier, and C. Pelachaud. Facial feedback signals for ECAs. In *AISB'07 Annual convention, workshop \Mindful Environments"*, pages 147{153, Newcastle upon Tyne, UK, April 2007.

(Bevacqua et al., 2010) E. Bevacqua, S. Chandra Pammi, M. Schröder, C. Pelachaud: Multimodal backchannels for Embodied Conversational Agents. International conference on Intelligent virtual agents IVA'10, Philadelphia, Sept 2010.

(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)

(Chartrand et al., 2005) Chartrand, T.L., Maddux, W. and Lakin, J.: Beyond the perception-behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry. In The new unconscious, edited by R. Hassin, J. Uleman, & J.A. Bargh, pp. 334-361, New York: Oxford University Press. (2005)

(de Sevin et al., 2010) E. de Sevin, S. Hyniewska, C. Pelachaud: Real-time Backchannel Selection for ECAs according to Personality. International conference on Intelligent virtual agents IVA'10, Philadelphia, Sept 2010.

(Eysenck, 1991) Eysenck, H. J.: Dimensions of personality: Criteria for a taxonomic paradigm. Personality and Individual Differences, 12, 773-79. (1991)

(Eysenck and Eysenck, 1978) Eysenck, S.B.G and Eysenck, H.J.: Impulsiveness and Venturesomeness - Their Position in a Dimensional System of Personality Description. Psychol Rep, 43(3):1247-1255 (1978)

(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)

(Heylen et al., 2007) D. Heylen, E. Bevacqua, M. Tellier, and C. Pelachaud. Searching for prototypical facial feedback signals. In *7th International Conference on Intelligent Virtual Agents*, pages 147{153, Paris, France, 2007.

(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.

(Lakin et al., 2003) Lakin, J. L., Jefferis, V. A., Cheng, C. M., and Chartrand, T. L. (2003). Chameleon effect as social glue: Evidence for the evolutionary significance of nonconsious mimicry. *Nonverbal Behavior*, 27(3):145{162.

(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, 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.

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

(Niewiadomski et al., 2009) 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.

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

(Peters et al., 2005) Peters, C, Pelachaud, C, Bevacqua, E, Poggi, I, Mancini, M, and Chafai, N. E. A model of attention and interest using gaze behavior. In *Proceeding of IVA'05: Intelligent Virtual Agents*, Kos, Greece, September 12-14 2005.

(Poggi, 2003) Poggi, I. (2003). Mind markers. In Trigo, N., Rector, M., and Poggi, I., editors, *Gestures. Meaning and use*. University Fernando Pessoa Press, Oporto, Portugal.

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

(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.