

# The NECA Project: Net Environments for Embodied Emotional Conversational Agents

## Project Note

Brigitte Krenn  
Austrian Research Institute for Artificial Intelligence ÖFAI  
Vienna, Austria  
[brigitte@oefai.at](mailto:brigitte@oefai.at)  
<http://www.oefai.at/NECA/>

### Introduction

The purpose of the NECA project (<http://www.oefai.at/NECA/>) is to develop a platform for the implementation of emotional conversational agents for Web-based applications. The users watch embodied conversational agents engage in verbal and nonverbal interaction in virtual “locations” on the internet. In the demonstrators, eShowroom and Socialite, two such “locations” are realised.

While in the first project year the core techniques, programs and resources to enable the generation of animated conversation in web-based applications have been developed and implemented, the second project year is dedicated to the refinement of the system modules and the enhancement of the functionality of the demonstrators, with a focus on affective reasoning, and verbal and nonverbal realization of expressive behaviour. The runtime of the project is 30 months.

In the following we give an introduction to the demonstrators, followed by an overview of the NECA architecture, and then describe aspects of emotion modelling in NECA.

### The Demonstrators

In the **eShowroom** scenario a car sales dialogue between a seller and a buyer is simulated. The purpose of this application is to entertain the site visitor and to embed product information into a narrative context similar to TV commercials nowadays. User interaction is restricted to setting general parameters prior to the display. These parameters include the user’s preferences in respect to different value dimensions, e.g. on how important aspects like sportiness, prestige or environmental issues are for the user. After having specified these preferences, a scene is generated which takes these settings into account: The agents/interlocutors will put special emphasis on conveying information about those aspects, which have been classified as being of importance for the user. The user can also specify the personality traits of the agents, i.e., their agreeableness and politeness, to influence the style

and the course of their conversation. This option aims to provide a means of entertaining the user by experimenting with different (possibly absurd) settings.



Figure 1: Screenshot eShowroom

A screenshot of a web browser window displaying the NECA Flash Player interface. The browser title is "NECA Flash Player - Microsoft Internet Explorer". The page header includes the website "derSpittelberg.at" and the NECA logo with the tagline "NET ENVIRONMENT FOR EMBODIED EMOTIONAL CONVERSATIONAL AGENTS". The interface is divided into several sections: a "NEXT EVENT" section with the text "Noch nichts im Plan.", a calendar for "OKT/NOV" showing the date "29", and a list of events for "29.10.2002". The main content area features a 3D scene of two avatars in a bar setting. Below the scene is a chat window with the following text: "Dienstag 01:36 an der lex-Bar", "I met Heike yesterday - do you know her?", "Sure, she is really nice.", "Well...", "That's not the impression I got!", and "What a pain...". The interface also includes a "SCORE: 194" and "STIMMUNG" meter, and a "PLAY AGAIN" button. The footer contains navigation links like "INFO", "TELL A FRIEND", "HIGHSCORE", "CHAT", "LOGOUT", and "created by sysis".

## Figure 2: Screenshot: Socialite<sup>1</sup>

The **Socialite** demonstrator implements a multi user web-application in the social domain. The users create their personal avatar, endow it with personality traits and preferences and send it to the virtual environment in order to meet other avatars. The overall goal is to be accepted in the community, to reach a certain degree of popularity within this environment. In this setting the user is not permanently logged on. The avatar/agent will report back to the user about encounters with other avatars when the user logs in the next time. This report is presented in the form of textual monologues, which are alternated with displays of animated dialogues between avatars. The user is then queried for choosing new instructions for her avatar from a given set of possibilities and sends the avatar off to its environment again. Socialite is part of a Web community called derSpittelberg which stands for a community of flat sharing students living in a particular area of Vienna named Spittelberg. The demonstrators can be accessed via the NECA homepage.<sup>2</sup>

## NECA Architecture and Rich Representation Language

The NECA architecture consists of the following main components: a scene generator, an affective reasoning component, a multimodal natural language generator (MNLG), a text/concept-to-speech synthesis, and a gesture assignment module. The information available at the interfaces between the system components is encoded in an abstract XML-compliant representation scheme which all together constitutes the NECA RRL (Rich Representation Language). The interfaces from the gesture assignment module to the animation engines are mappings from the player unspecific representations in RRL to player specific code. The NECA architecture is described in more detail in [Krenn et al., 2002].

The *Rich Representation Language* has been designed for the description of agent behaviour in our net environments. The RRL represents a wide range of expert knowledge required at the interfaces between the different components in the NECA architecture. Thus the RRL differs from multimedia markup languages such as MPML or TVML<sup>3</sup> which are typically designed to support a strongly text-based annotation of multimodal input to media players in a rather generalized way. The NECA RRL is much more comparable to behaviour generation languages which have been created in the context of a certain system architecture, its components and a particular application scenario. See for instance APMML which is part of the MagiCster system architecture and is used for the markup of information required for behaviour generation of a talking head in a medical diagnosing situation, [De Carolis et al, 2002]. This however does not mean that there are no commonalities between these representation languages. In developing the RRL we draw as much as possible on existing

<sup>1</sup> The screenshot is taken from a demonstrator for an international audience therefore the text displayed below the animation window is an English translation of the German spoken dialogue. In the online version, the German text is displayed.

<sup>2</sup> [www.oefai.at/NECA/](http://www.oefai.at/NECA/)

<sup>3</sup> <http://www.miv.t.u-tokyo.ac.jp/HomePageEng.html>,  
<http://www.strl.nhk.or.jp/TVML/index.html>

standardisation efforts and approaches to XML-based behaviour markup. For a description of the RRL see [Piwek et al., 2002]. The full specification of the RRL can be obtained from <http://www.oefai.at/NECA/RRL/index.html>.

In the preparation phase of the eShowroom application, the user is able to provide the system with information on her information needs and preferences. This will be used as a basis for the *scene generator*. In the case of Socialite, the user on first login to the application domain ,derSpittelberg, defines personality traits, interests and outer appearance of her avatar. In addition the user is able to influence the avatar during the runtime of the application. The history of the avatar's existence/encounters in the application and the influence the user exerts on the avatar have an effect on the scene generation in Socialite. In both demonstrators the scene generator takes the role of a playwright, generating a script for the characters that become actors in a scene. In the script, the communicative acts to be carried out are specified as well as their temporal coordination and the emotion associated.

The *multimodal natural language generator* transforms the formal specification of the communicative acts into text, annotated with syntactic, semantic, and pragmatic features. The component is also responsible for selecting meaning-bearing and (discourse) functional gestures. Even though multimodal natural language generation is application dependent, we aim at a common core technology which can be applied in both demonstrators, see [Piwek, 2003]. A crucial aspect here is that the generation component supports template generation as well as full generation. Template based generation is called for when the application requires generation of colloquial utterances which reflect a certain sociolect as it is the case in the Socialite scenario. The eShowroom scenario on the other hand is more appropriate for the use of full generation, as a major aspect here is the generation of information on particular features of a car which have been identified as important for the buyer.

The task of the *text/concept-to-speech synthesis* is then to convey, through adequate voice quality and prosody, the intended meaning of the text as well as the emotion with which it is uttered. It also provides information on the exact timing of utterances, syllables and phonemes, which is indispensable for synchronisation of verbal and nonverbal aspects of behaviour in the gesture assignment module. To achieve these goals we use MARY, a concatenative synthesis system, see [Schröder, Trouvain, to appear], together with expressive voices which have been created within the project. As MARY is a German synthesis system it can only be integrated into the Socialite application. For eShowroom, where English utterances are generated, we have to rely on systems that are available outside the consortium. Here we have to cope with the problem that current synthesis components typically do not output timing information, not to speak about the non-availability of expressive voices.

The *gesture assignment* module builds on the information regarding facial expression and content-bearing gestures available from MNLG, on the emotion type and intensity available from affective reasoning, and on the timing information available after speech synthesis. The task of the gesture assignment component is to select appropriate gestures and facial animations from a gesture lexicon (gesticon) and to schedule the nonverbal behaviours in accordance with the voice/spoken utterance. This requires the ability of the assignment module to concatenate gestures, to speedup and prolong gestures, to synchronize the speed of the gestures with the speech rate, to align gestures or parts of gestures to a certain part of the utterance such as a phrase, a word or in the case of facial expressions to align eye brow raises to certain accented syllables, eye gaze to turns in the dialogue, etc. This functionality is currently under implementation. The gesture assignment module is also responsible for the synchronization of phonemes and visemes. The gesticon is XML compliant. Gesture entries

are specified with respect to (i) spacio-temporal characteristics, i.e., the spacial orientation of the hand at the beginning and the end of a stroke, and the default, minimal and maximal duration of the stroke, (ii) semantic and functional information, i.e., whether a gesture is iconic, deictic, emblematic or a beat; where it should be attached to, e.g. to a certain dialogue act, a phrase, word, etc.; how it should be aligned: sequential or parallel, parallel to the beginning of or the end of the element the gesture is attached to; what is the meaning of the gesture, (iii) a link to the player specific code, (iv) restrictions as regards the applicability of the gesture with respect to emotion, personality and activity. As regards the spacio-temporal markup of gestures we have drawn on MURML [Kranstedt et al., 2002]. For NECA purposes, however, we have reduced the levels and the granularity of the description in MURML.

The gesticon entries for facial expressions comprise information on (i) restrictions, (ii) the meaning or function of a facial expression and what the signal is, e.g. smile, frown, brows up, etc., (iii) player specific code. (ii) relates to the “communicative functions” in APML which are defined as meaning-signal pairs such as <joy,smile> or <backchannel,smile>, see [De Carolis et al., 2002].

The output of gesture assignment is a *control sequence* comprising the synchronised verbal and nonverbal behaviour of all the characters in the scene. In an extra processing step this control sequence is converted into a data stream that can be processed by a specific player. Up to date we have worked with Microsoft Agent ([www.microsoft.com/msagent/default.asp](http://www.microsoft.com/msagent/default.asp)) and Charamel ([www.charamel.de](http://www.charamel.de)) for eShowroom and Macromedia Flash ([www.macromedia.com](http://www.macromedia.com)) for Socialite. Depending on the player technology used, the granularity of the integration of verbal and nonverbal behaviour is influenced. In Microsoft Agents for instance parallel or overlapping rendering of gesture and speech is impossible. In Flash, to the other extreme, integration of verbal and nonverbal behaviour can be as fine-grained as the alignment of an eyebrow raise with a phoneme. This, however, requires a certain proficiency in Flash programming. Charamel also offers a greater flexibility than Microsoft Agent and an advantage over Flash is that it is scriptable, and therefore there is less programming effort involved. A disadvantage however is, that Charamel is a proprietary format whereas in Flash the developers of an application have full control over the design and the creation of animations. In the beta version of the NECA demonstrators, Flash and Charamel animation will be incorporated. Microsoft Agent has been applied as a minimal approach in the first version of the eShowroom demonstrator, whereas in the first Socialite demonstrator the feasibility of Flash for the task at hand has been explored and demonstrated.

## **Aspects of Emotion Modelling in NECA**

Within the NECA project we make use of two approaches to emotion modelling. One is based on the OCC model [Ortony, Clore, Collins, 1988] and is applied for specifying the emotion types which are assigned to communicative acts. The other approach is based on emotion dimensions: activation, evaluation, power (see Cowie et al., 1999) and is used in speech synthesis with the MARY system in order to express shades of emotions in the voice quality, and thus allows changes of emotional tone to be conveyed over time.

With regard to emotion types assigned to communicative acts in the simulated dialogues, we have made provisions in the definition of the RRL to distinguish between the emotions which the speaker character feels when performing the act and the ones which the character expresses in performing the act. Additionally, both the emotion felt by the hearer character when the act is performed and the emotions expressed as a direct result of the dialogue act are part of the (RRL) description of a communicative act. This distinction becomes relevant when

a model of display rules is integrated into the NECA system. This is work which has not been tackled yet. At the current state each dialogue act in the scene is assigned with the emotion felt. This is done by the EmotionEngine, an implementation of the OCC model, [Gebhard et al., to appear]. The emotion assigned influences the lexical realization of the utterance -- for instance the lexical selection of adjectives --, the tone of voice, and the generation of facial expressions accompanying the act. Based on the work described in [Cowie et al., 1999] a mapping between the OCC emotion types and the emotion dimensions used in speech synthesis has been developed. In addition to the emotion type, an intensity value for the emotion is specified by the EmotionEngine, this value is then used in the gesture assignment module to control the intensity of gestures and the physiologically motivated characteristics of the animation such as the intensity of eye blinks and breathing.

In speech synthesis, the NECA project introduces an innovative approach to making synthesis flexible and expressive, by recording diphone voices with three phonetically defined voice qualities (neutral, soft, loud), and by explicitly modelling the acoustic properties of emotions described in terms of emotion dimensions, see [Schröder et al., 2001]. This is possible because of the rich segmental material, and its combination with the flexible control of prosodic features in the diphone synthesis technology employed in MARY. The implementation of a mapping from emotion dimensions to acoustic properties of the synthetic voice results in unprecedented flexibility in emotion expression through synthetic speech, see [Schröder, Grice, to appear].

## **Acknowledgements**

The work described here is joint work of the members of the NECA consortium, in particular Brigitte Krenn, Hannes Pirker and Neil Tipper of the Austrian Research Institute for Artificial Intelligence (ÖFAI), Marc Schröder of the Language Technology Group of the German Research Institute for Artificial Intelligence (DFKI), Martin Klesen and Thomas Rist of DFKI's Intelligent User Interfaces Group, Stefan Baumann and Martine Grice of the Department of Phonetics of the Saarland University, Kees van Deemter and Paul Piwek of the Information Technology Research Institute (ITRI) of the University of Brighton, and Erich Gstrein, Bernhard Herzog, Markus Bruckner and Barbara Neumayr of Sysis, a Vienna based company which develops amongst others community-building applications for the Internet. Peter Rist has designed the characters for the eShowroom and Jan Wachter the characters for Socialite. Cordula Klein and Uta Panten, both students at the Phonetics Department of Saarland University have worked on the labelling of the data bases for expressive German voice. This research is supported by the EC Project NECA IST-2000-28580.<sup>4</sup> ÖFAI is supported by the Austrian Federal Ministry for Education, Science and Culture.

## **Bibliography**

[Cowie et al., 1999] Cowie, R., Douglas-Cowie, E., Appolloni, B., Taylor, J., Romano, A., Fellenz, W. (1999). What a neural net needs to know about emotion

<sup>4</sup> Disclaimer: The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

words. In Mastorakis, N. (ed.). Computational Intelligence and Applications. World Scientific & Engineering Society Press.

[De Carolis et al, 2002]. De Carolis, B., Carofiglio, V., Bilvi, M., Pelachaud, C. (2002) APML, a Mark-up Language for Believable Behaviour Generation. In *Proceedings of the Workshop ``Embodied conversational agents - let's specify and evaluate them!''*, held in conjunction with AAMAS-02, July 16 2002, Bologna, Italy.

[Gebhard et al., to appear] Gebhard, P., Kipp M., Klesen M., Rist, T. (to appear). Adding the Emotional Dimension to Scripting Character Dialogues. to appear In *Proceedings of the 4th International Working Conference on Intelligent Virtual Agents (IVA'03)*, Kloster Irsee, 2003.

[Kranstedt et al., 2002] Kranstedt, A., Kopp, St., Wachsmuth, I. (2002). MURML: A Multimodal Utterance Representaiton Markup Language for Conversational Agents. In *Proceedings of the Workshop ``Embodied conversational agents - let's specify and evaluate them!''*, held in conjunction with AAMAS-02, July 16 2002, Bologna, Italy.

[Krenn et al., 2002] Krenn, B., Grice, M., Piwek, P., Schröder, M., Klesen, M., Baumann, S., Pirker, H., van Deemter, K., Gstrein, E. (2002). Generation of Multimodal Dialogue for Net Environment. In *Proceedings of KONVENS-02*, 30 September - 2 October 2002, Saarbrücken, Germany.

[Ortony, Clore, Collins, 1988] Ortony, A., Clore, G. and Collins, A. (1988). *The Structure of Emotions*. Cambridge University Press. Cambridge MA.

[Piwek, 2003] Piwek, P. (2003). A Flexible Pragmatics-driven Language Generator for Animated Agents. In: *Proceedings of EACL03 (Research Notes Companion Volume)*, Budapest, 151-154.

[Piwek et al., 2002] Piwek, P., Krenn, B., Schröder, M., Grice, M., Baumann, S., Pirker, H. (2002). RRL: A Rich Representation Language for the Description of Agent Behaviour in NECA. In *Proceedings of the Workshop ``Embodied conversational agents - let's specify and evaluate them!''*, held in conjunction with AAMAS-02, July 16 2002, Bologna, Italy.

[Schröder et al., 2001] Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M. and Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Proceedings of Eurospeech*, Aalborg, Denmark. Volume 1, 87-90.

[Schröder, Grice, to appear] Schröder, M. and Grice, M. (to appear). Expressing vocal effort in concatenative synthesis. to appear In *15th International Conference of Phonetic Sciences*, Barcelona, Spain.

[Schröder, Trouvain, to appear] Schröder M. and Trouvain J. (to appear). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. to appear In *International Journal on Speech Technology*, Special Issue following the 4th ISCA Workshop on Speech Synthesis.

