2009 IEEE International Conference on Robotics and Automation
Kobe International Conference Center
Kobe, Japan, May 12-17, 2009

Generating Robot/Agent Backchannels During a Storytelling Experiment

S. Al Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, A. Mahdhaoui,
J.-C. Martin, S. Ondas, C. Pelachaud, J. Urbain, M. Yilmaz

Abstract— **This work presents the development of a real-time framework for the research of Multimodal Feedback of Robots/Talking Agents in the context of Human Robot Interaction (HRI) and Human Computer Interaction (HCI). For evaluating the framework, a Multimodal corpus is built (ENTERFACE_STEAD), and a study on the important multimodal features was done for building an active Robot/Agent listener of a storytelling experience with Humans. The experiments show that even when building the same reactive behavior models for Robot and Talking Agents, the interpretation and the realization of the behavior communicated is different due to the different communicative channels Robots/Agents offer be it physical but less human-like in Robots, and virtual but more expressive and human-like in Talking agents.**

I. INTRODUCTION

During the last years, several methods have been proposed for the improvement of the interaction between humans and talking agents or robots. The key idea of their design is to develop agents/robots with various capabilities: establish/maintain interaction, show/perceive emotions, dialog, display communicative gesture and gaze, exhibit distinctive personality or learn/develop social capabilities [1], [2]. Theses social agents and robots aim at naturally interacting with humans by the exploitation of these capabilities. In this paper, we have investigated one aspect of this social interaction: the engagement in the conversation [3]. The engagement process makes it possible to regulate the interaction between the human and the agent or the robot. This process is obviously multi-modal (verbal and non-verbal) and requires an involvement of both the partners.

This paper deals with two different interaction types namely Human-Robot Interaction (HRI) with the Sony AIBO robot and Human-Computer Interaction (HCI) with an Embodied Conversational Agent (ECA). The term ECA has been coined in Cassell et al. [4] and refers to human-like virtual characters that typically engage in face-to-face communication with the human user. We have used GRETA [5], an ECA, whose interface obeys the SAIBA (Situation, Agent, Intention, Behavior, Animation) architecture [6]. We focused on the design of an open-source, real-time software platform for designing the feedbacks provided by the robot and the humanoid during the interaction[1]. The multimodal feedback problem we considered here was limited to facial and neck movements by the agent (while the AIBO robot uses all possible body movements, given its poor facial expressivity): we did not pay attention to arms or body gestures.

This paper is organized as follows. In section II, we present the storytelling experiment used for the design of our human robot/agent interaction system described in section III. Section IV focuses on speech and face analysis modules we have developed. We then give in sections V and VI a description of the multi-modal generation of backchannels including interpretation of communicative signals and the implemented reactive behaviors of the agent and the robot. Finally, section VII presents the details of the evaluation and comparison in our HCI and HRI systems.

II. FACE-TO-FACE STORYTELLING EXPERIMENT

A. Data collection

In order to model the interaction between the speaker and the listener during a storytelling experiment, we first recorded and annotated a database of human-human interaction termed eNTERFACE_STEAD. This database was used for extracting feedback rules (section II-B) but also for testing the multimodal feature extraction system (section IV).

We followed the McNeill lab framework [7]: one participant (the speaker), has previously observed an animated cartoon (Sylvester and Tweety), retells the story to a listener immediately. The narration is accompanied by spontaneous communicative signals (filled pauses, gestures, facial expressions...). 22 storytelling sessions were videotaped with different conditions: 4 languages (Arabic, French, Turkish and Slovak). The videos have been annotated (with at least two annotators per session) for describing simple communicative signals of both speaker and listener: smile, head nod, head shake, eye brow and acoustic prominence.

S. Al Moubayed is with Center for Speech Technology, Royal Institute of Technology KTH, SWEDEN sameram@kth.se
M. Baklouti is with the Thalès, FRANCE malek.baklouti@thalesgroup.com
M. Chetouani and A. Mahdhaoui are with the University Pierre and Marie Curie, FRANCE mohamed.chetouani@upmc.fr, Ammar.Mahdhaoui@isir.fr
T. Dutoit and J. Urbain are with the Faculté Polytechnique de Mons, BELGIUM, thierry.dutoit@fpms.ac.be, jerome.urbain@fpms.ac.be
J.-C. Martin is with the LIMSI, FRANCE martin@limsi.fr
S. Ondas is with the Technical University of Kosice, SLOVAKIA stanislav.ondas@gmail.com
C. Pelachaud is with the INRIA, FRANCE catherine.pelachaud@inria.fr
M. Yilmaz is with the Koc University, TURKEY yilmazmehmetmustafa@gmail.com

[1]The database and the source code for the software developed during the project are available online from the eNTERFACE08 web site: www.enterface.net/enterface08.

TABLE I
AGREEMENT AMONG ANNOTATORS

Track name	Agreement (%)
Speaker_Face	89.3
Speaker_Acoustic	84.5
Listener_Face	77.96
Listener_Acoustic	95.97

Manual annotations of videos were evaluated by computing agreements using corrected kappa [8] computed in the Anvil tool [9]. Table I presents the agreements among annotators for each track. We can see that the best agreement is obtained for the Listener_Acoustic track which is expected since the listener is not assumed to speak and when he/she does simple sounds are produced (filled pauses). Other tracks have a lower agreement such as Speaker_Acoustic. The speaker always speaks during the session and prominent events are less identifiable. However, the agreements measures are high enough to allow us to assume that selected communicative signals might be reliably detected.

B. Extracting rules from data

Based on the selected communicative signals, we have defined some rules to trigger feedbacks. The rules are based on [10], [11], which involved mainly only mono-modal signals. The structure of such rules is as follows:

If some *signal* (eg. head-nod — pause — pitch accent) is received, then the listener sends some *feedback_signal* with probability X.

We have extended these rules by analyzing the data annotated from our storytelling database. We looked at correlation between, not only, speakers mono-modal signal and listeners feedback, but also we studied the relation between speakers multi-modal signals and feedback. We define multi-modal signals as any set of overlapping signals that are emitted by the speaker.

For each mono-modal (resp. multi-modal) signal emitted by the speaker we calculate their number of occurrences. Within the time-window of each speakers signal, we look at co-occurring listeners signals. We compute the correlation of occurrence between each speakers signal and each listeners signal. This computation gives us a correlation matrix between speakers and listeners signals. This matrix can be interpreted as: given a speakers signal, the probability that the listener would send a given signal. In our system we use this matrix to select listeners feedback signals. When a speakers signal is detected, we choose from the correlation matrix, the signal (i.e. feedback) with the higher probability.

From this process, we identified a set of rules[2], among them:

- Mono-modal signal ⇒ mono-modal feedback: *head_nod* is received, then the listener sends *head_nod_medium*.
- Mono-modal signal ⇒ multi-modal feedback: *smile* is received, then the listener sends *head_nod and smile*.
- Multi-modal signal ⇒ mono-modal feedback: *head_activity_high and pitch_prominence* are received, then the listener sends *head_nod_fast*.
- Multi-modal signal ⇒ multi-modal feedback: *pitch_prominence and smile* are received, then the listener sends *head_nod and smile*.

These rules are implemented in our system in order to trigger feedbacks, the multi-modal fusion module makes it possible to activate these rules (section V).

[2]A complete list can be found at: http://www.enterface.net/enterface08

III. SYSTEM DESIGN

Although Human beings are all perfectly able to provide natural feedback to a speaker telling a story, explaining how and when you do it is a complex problem. ECAs are increasingly used in this context to study and model human-human communication as well as for performing specific automatic communication tasks with humans.

Examples are REA [12], an early system that realizes the full action-reaction cycle of communication by interpreting multimodal user input and generating multimodal agent behavior. Gandalf [13] provides real-time feedback to a human user based on acoustic and visual analysis. In robotics, various models have been proposed for the integration of feedbacks during interaction [2]. Recently, the importance of feedbacks for discourse adaptation has been highlighted during an interaction with BIRON [14].

In a conversation, all interactants are active. Listeners provide information to the speaker their view and engagement in the conversation. By sending acoustic or visual feedback signals, listeners show if they are paying attention, understanding or agreeing with what is being said. Taxonomies of feedbacks, based on their meanings, have been proposed [15], [16]. The key idea of this project is to automatically detect the communicative signals in order to produce a feedback. Contrary to the approach proposed in [14], we focus on non-linguistic features (prosody, prominence) but also on head features (activity, shake, nod).

Our system is based on the architecture proposed by [5], but progressively adapted to the context of a storytelling (figure 1). We developed several modules for the detection and the fusion of the communicative signals from both audio and video analysis. If these communicative signals match our pre-defined rules, a feedback is triggered by the Realtime BackChannelling module resulting on two different messages (described in section VI) conveying the same meaning.

IV. MULTI-MODAL FEATURE EXTRACTION

A. Speech Analysis

The main goal of the speech analysis component is to extract features from the speech signal that have been previously identified as key moments for triggering feedbacks (cf. section II). In this study, we do not use any linguistic information to analyze the meaning of the utterances being

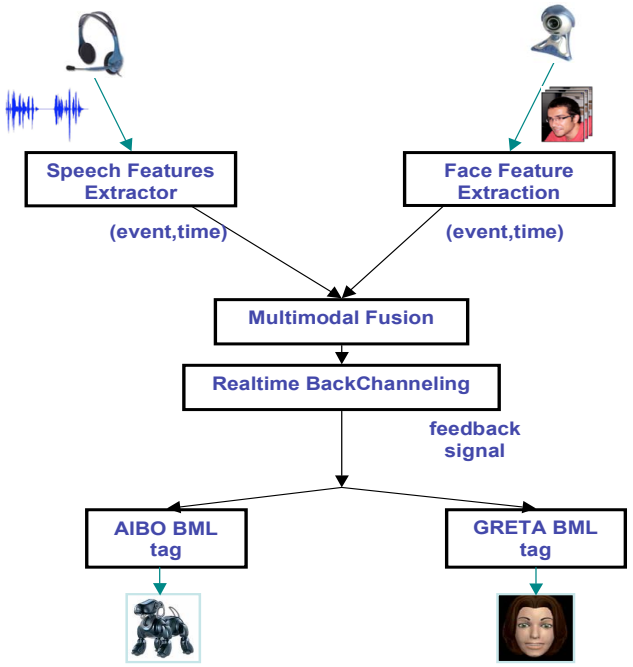

Fig. 1. Architecture of our interaction feedback model.

told by the speaker, but we focus on the prosodic cross-language features which may participate in the generation of the feedback by the listener.

1) Feature Extraction: Previous studies have shown that pitch movements, specially at the end of the utterances, play an important role in turn taking and backchannelling during human dialogue [10]. We propose in this work to use the following features extracted from the speaker's speech signal: Utterance beginning, Utterance end, Raising pitch, Falling pitch, Connection pitch, and Pitch prominence (cf. section II).

To extract utterances beginning and ending, a realtime implementation of a Voice Activity Detector (VAD), which is an adaptation of the SPHINX Vader functionality [18], has been developed. To extract pitch movements, we used an implementation of the realtime fundamental frequency tracking algorithm YIN [19]. We compensate outliers and octave jumps of the F0 by a median filter of size 5 (60 msec). After extracting the pitch, the TILT model [20] is used to extract Raising pitch, Falling pitch and Connection pitch.

These algorithms are then used as a package in PureData (PD)[17], a graphical programming environment for real-time audio processing, PD is used as an audio provider with 16KHz audio sampling rate. This package sends the id of the features in the speech signal to the multi-modal fusion model whenever any of these features is detected.

2) Pitch Prominence Detection: In the literature, several definitions of acoustical prominent events can be found showing the diversity of this notion [21], [22]. Terken [22] defines prominence as words or syllables that are perceived as standing out from their environment. Most of the proposed definitions are based on linguistic and/or phonetic units.

We propose, in this paper, another approach using statistical models for the detection of prominence. The key idea is to assume that a prominent sound stands out from the previous message. For instance, during our storytelling experiment, speakers emphasize words, syllables when they want to focus the attention of the listener on important information. These emphasized segments are assumed to stand out from the overall ones, which makes them salient.

Prominent detectors are usually based on acoustic parameters (fundamental frequency, energy, duration, spectral intensity) and machine learning techniques (Gaussian Mixture Models, Conditional Random Fields)[23], [24]. Unsupervised methods have been also investigated such as the use of Kullback-Leibler (KL) divergence as a measure of discrimination between prominent and non-prominent classes [25]. These statistical methods provide an unsupervised framework adapted to our task. The KL divergence needs the estimation of two covariance matrices (Gaussian assumption):

$$KL_{ij} = \frac{1}{2}[\log\frac{\Sigma_j}{\Sigma_i} + \mathrm{tr}(\Sigma_i\Sigma_j^{-1}) + (\mu_i-\mu_j)^T\Sigma_j^{-1}(\mu_i-\mu_j) - d] \quad (1)$$

where μ_i, μ_i and Σ_i, Σ_j denote the means and the covariance matrices of i-th (past) and j-th (new event) speech segments respectively. d is the dimension of the speech feature vector. An event j is defined as prominent if the distance from the past segments (represented by the segment i) is larger than a pre-defined threshold.

One major drawback of the KL divergence approach is that since the new event is usually shorter, in terms of duration, than the past events, the estimation of covariance matrices is less reliable. In addition, it is well-known that duration is an important perceptual effect for the discrimination between sounds. Taking into account these points, we propose to use another statistical test namely the T^2 Hotteling distance defined by:

$$H_{ij} = \frac{L_iL_j}{L_i+L_j}[(\mu_i-\mu_j)^T\Sigma_{i\cup j}^{-1}(\mu_i-\mu_j)] \quad (2)$$

where $i \cup j$ is the union of i-th (past) and j-th (new event) segments. L_i and L_j denote the length of the segments. T^2 Hotteling divergence is closely related to the Mahalanobis distance.

In this work only the fundamental frequency (F0) is used as a feature to calculate the Hotteling distance between two successive voiced segments. In this sense, a prominence is detected when the Hotelling distance between the current and the preceding Gaussian distributions of F0 is higher than a threshold. The decision is done by the help of a decaying distance threshold over time: adaptation to the speaker. Since we estimate a statistical model of the pitch for a voiced segment, we only estimate it when there is enough pitch samples during the voiced segment, set to 175 msec.

B. Face Analysis

The main goal of the face analysis component (figure 2) is to provide the feedback system with some knowledge of

communicative signals conveyed by the head of the speaker. More specifically, detecting if the speaker is shaking the head, smiling or showing neutral expression are the main activity features we are interested in. The components of this module are responsible for face detection, head shake and nod detection, mouth extraction, and head activity analysis. They are detailed below.

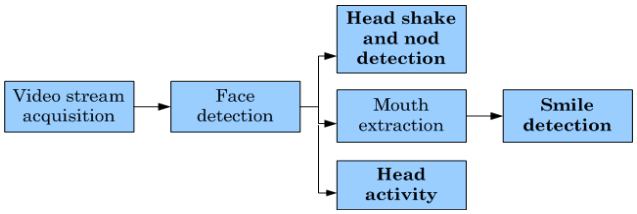

Fig. 2. Overview of the face analysis module.

1) Face Detection: The face detection algorithm that we used exploits Haar-like features that have been initially proposed by Viola & Jones [26]. It is based on a cascade of boosted classifiers working with Haar-like features and trained with a few hundreds of sample views of faces. We used the trained classifier available in OpenCV. The face detection module outputs the coordinates of existing faces in the incoming images.

2) Smile Detection: Smile detection is performed in two steps: mouth extraction followed by smile detection. We use a colorimetric approach for mouth extraction. A thresholding technique is used after a color space conversion to the YIQ space. Once the mouth is extracted, we examine the ratio between the two characteristic mouth dimensions, P_1P_3 and P_2P_4 (figure 3), for smile detection. We assume that when smiling, this ratio increases. The decision is obtained by thresholding.

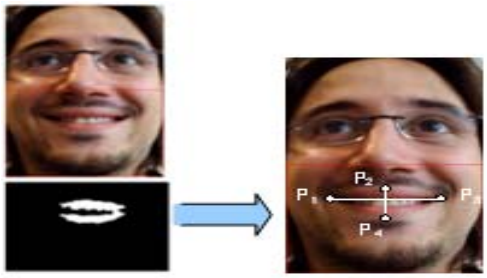

Fig. 3. Smile detection: combining colorimetric and geometric approaches.

3) Head shake and nod detections: The purpose of this component is to detect if the person is shaking or nodding the head. The idea is to analyze the motion of some feature points extracted from the face along the vertical and horizontal axes. Once the face has been detected in the image, we extract 100 feature points using a combined corner and edge detector defined by Harris [27]. Feature points are extracted in the central area of the face rectangle using offsets. These points are then tracked by calculating the optical flow between a set of corresponding points in two successive frames. We make use of the Lucas-Kanade [28] algorithm implementation available in the OpenCV library.

Let n be the number of feature points and $Pt_i(x_i, y_i)$ the i-th feature point defined by its 2D screen coordinates (x_i, y_i). We then define the overall velocity of the head as:

$$V = \begin{cases} V_x &= \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{i-1}) \\ V_y &= \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{i-1}) \end{cases} \quad (3)$$

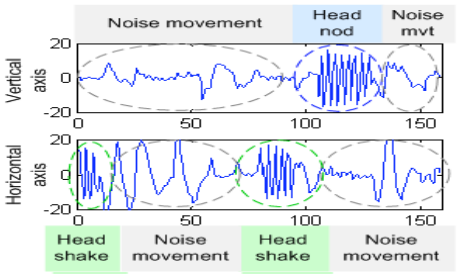

Fig. 4. Feature point velocity analysis.

Figure 4 shows the velocity curves along the vertical and horizontal axes. The sequence of movements represented is composed by one nod and two head shakes. We notice that the velocity curves are the sum of two signals: (1) a noise movement which is a low frequency signal representing the global head motion and (2) a high frequency signal representing the head nods and head shakes.

The idea is then to use wavelet decomposition to remove the low frequency signals. More precisely, we decomposed the signal using symlet-6 wavelet. Figure 5 shows the reconstruction of the details at the first level of the signal shown in figure 4. The head nod and shake events can be reliably identified by this process.

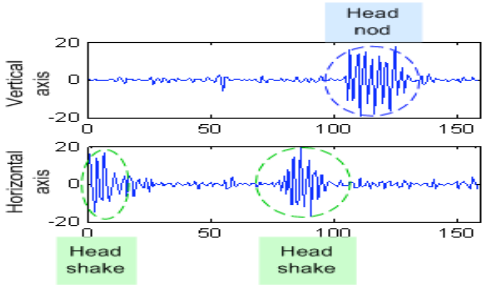

Fig. 5. Signal denoising via wavelets.

4) Head activity analysis: Analysis of recordings of the storytelling experience has shown a correlation between the head activity of both speaker and listener. To characterize the head activity, we use the velocity of the feature points, defined in (3), to quantify the overall activity A:

$$A = \sum_{i \in time\_window} V_{x,t}^2 + V_{y,t}^2 \quad (4)$$

where the $time\_window$ is set to 60 frames (30 frames/s)

TABLE II
QUANTIZATION OF THE HEAD ACTIVITY

Amplitude	Interpretation
$<$ mean	LOW ACTIVITY
$<$ mean + standard deviation	MEDIUM ACTIVITY
Otherwise	HIGH ACTIVITY

This measure provides information about the head activity levels. In order to quantize head activity into levels (high, medium or low), we analyzed the head activity of all the speakers of the eNTERFACE_STEAD corpus. Assuming that the activity of one given speaker is Gaussian, we set up different thresholds defined in table II. By using these thresholds, the algorithm will become more sensitive to any head movement of a stationary speaker whereas it will raise the thresholds for an active speaker resulting on a flexible adaptive modeling.

V. MULTI-MODAL FUSION

The Multimodal Fusion Module works by the principle of activating probabilistic rules (cf. section II-B) depending on the multimodal events it receives. When a rule is completed then the output of the rule is sent as a message to the different Agents/Robots connected to it as a feedback signal.

The rules in this work are extracted from the analysis of a database annotations and hand-written using feedback rules defined in the literature [10] (section II-B). The rule takes a list of input events (mono or multi modal) as output the rules defines one output feedback signal (mono or multi modal). The rule can be probabilistic by defining a probability of this rule, so in case there are more than one rule with the same input, every rule will have a probability of execution. For realtime consideration, the rule contains a response time variable, which defines when the output of the rule should be executed after the reception of the last input signal. If not all the input signals are received, the rule will be deactivated after this specified period.

VI. REACTIVE BEHAVIORS

In our architecture, we aim to drive simultaneously different types of virtual and/or physical agents (figure 1). To ensure high flexibility we are using the same control language to drive all the agents, the Behavior Markup Language BML [6]. BML encodes multimodal behaviors independently from the animation parameters of the agents.

Through a mapping we transform BML tags into MPEG-4 parameters for the GRETA agent and into mechanical movements for the AIBO robot. Various feedbacks are already available for GRETA such as acceptance (head_nod), non-acceptance (head_shake) or smile. Concerning AIBO, we developed similar feedbacks conveying the same meaning but in a different way. To develop the reactive behavior of AIBO, we used the URBI (Real-Time Behavior Interface) library [29] allowing a high-level control of the robot.

VII. ASSESSMENT AND DISCUSSION

A. Experimental setup

Evaluation research is still underway for virtual characters [30], [31] and for human-robot interaction [33]. Since the goal of the project was to compare feedback provided by two types of embodiments (a virtual character and a robot) rather than to evaluate the multi-modal feedback rules implemented in each of these systems, we decided to have users tell a story to both GRETA and AIBO at the same time. An instruction form was provided to the subject before the session. Then users watched the cartoon sequence, and were asked to tell the story to both AIBO and GRETA (figure 6). Finally, users had to answer a questionnaire. The questionnaire was designed to compare both systems with respect to the realization of feedback (general comparison between the two listeners, evaluation of feedback quality, perception of feedback signals and general comments). Sessions were videotaped using a Canon XM1 3CCD digital camcorder.

The current evaluation aims at evaluating the relevance of the characterization of communicative signals for the regulation of interaction. We performed here only a pretest and an anova is not possible because the number of subjects is too small (10 users). In addition, no hypotheses have been done on the expected results from questionnaires.

Fig. 6. The assessment set-up.

As illustrated by figure 7, 8 out of 10 users estimated that GRETA understood better the story than AIBO. Yet, 8 out of 10 users felt that AIBO looked more interested and liked the story more than GRETA did.

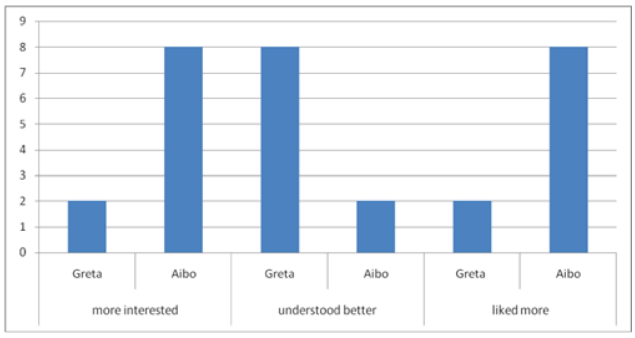

Fig. 7. Comparing the feedbacks provided by the virtual agent and the robot.

Further evaluations could be investigated with such a system. Another possibility would be to have the speaker tell two different stories one to GRETA, and then another one to AIBO. The order of the listeners should be counterbalanced across subjects. This would avoid having the speaker switching his attention between AIBO and GRETA. Perceptive tests on videos combining speakers and AIBO/GRETA listeners could also be designed to have subjects 1) compare random feedback with feedback generated by analyzing users behavior , or 2) rate if the listener has been designed to listen to this speaker or not.

VIII. CONCLUSIONS AND FUTURE WORKS

We presented a multi-modal framework to extract and identify Human communicative signals for the generation robot/agent feedbacks during storytelling. We exploited face-to-face interaction analysis by highlighting communicative rules. A real-time feature extraction module has been presented allowing the characterization of communicative events. These events are then interpreted by a fusion process for the generation of backchannel messages for both AIBO and GRETA. A simple evaluation was established, and results show that there is an obvious difference in the interpretation and realization of the communicative behavior between humans and agents/robots.

Our future works are devoted to the characterization of other communicative signals using the same modalities (speech and head). Prominence detection can be improved by the use of syllable-based analysis, which can be computed without linguistic information. Another important issue is to deal with the direction of gaze. This communicative signal conveys useful information during interaction and automatic analysis (human) and generation (robot/agent) should be investigated.

IX. ACKNOWLEDGMENTS

We are grateful to Elisabetta Bevacqua for her advice in the organization of our work and her help on interfacing our software with GRETA. We also want to acknowledge Yannis Stylianou for the feedback he gave during discussions on our project. This project was partly funded by Région Wallonne, in the framework of the NUMEDIART research program and by the FP6 IP project CALLAS.

REFERENCES

[1] T. Fong, I. Nourbakhsh and K. Dautenhahn, A Survey of Socially Interactive Robots, *Robotics and Autonomous Systems 42(3-4)*, 143-166, 2003.

[2] C. Breazeal, Social Interactions in HRI: The Robot View, R. Murphy and E. Rogers (eds.), *IEEE SMC Transactions*, Part C, 2004

[3] C.L. Sidner, C. Lee, C.D. Kidd, N. Lesh, C. Rich, Explorations in Engagement for Humans and Robots, *Artificial Intelligence*, May 2005

[4] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds). *Embodied Conversational Agents*. MIT Press, 2000.

[5] E. Bevacqua, M. Mancini, and C. Pelachaud, A listening agent exhibiting variable behaviour, *Intelligent Virtual Agents, IVA'08*, Tokyo, September 2008.

[6] H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, Z. Ruttkay, K. R. Thorisson, H. van Welbergen, R. van der Werf, The Behavior Markup Language: Recent Developments and Challenges, *Intelligent Virtual Agents, IVA'07*, Paris, September 2007.

[7] D. McNeil, Hand and mind: What gestures reveal about thought, Chicago IL, The University, 1992.

[8] R. L. Brennan, D. J. Prediger: Coefficient κ: Some uses, misuses, and alternatives. In: *Educational and Psychological Measurement*. 41,687699, 198.

[9] M. Kipp, Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, 1367-1370, 2001.

[10] R. M. Maatman, Jonathan Gratch, Stacy Marsella, Natural Behavior of a Listening Agent. *Intelligent Virtual Agents, IVA'05*, 25-36, 2005.

[11] N. Ward, W. Tsukahara, Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 23, 1177-1207, 2000.

[12] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjlmsson, H. Yan, Embodiment in Conversational Interfaces: Rea. *Proceedings of the CHI'99 Conference*, pp. 520-527. Pittsburgh, PA, 1999.

[13] J. Cassell and K. Thrisson, The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents, *Applied Artificial Intelligence*, 13(3), 1999.

[14] M. Lohse, K. J. Rohlfing, B. Wrede; G. Sagerer, "Try something else!" - When users change their discursive behavior in human-robot interaction, *IEEE Conference on Robotics and Automation*, Pasadena, CA, USA, 3481-3486, 2008.

[15] J. Allwood, J. Nivre, and E. Ahlsen. On the semantics and pragmatics of linguistic feedback. *Semantics*, 9(1), 1993.

[16] I. Poggi. Backchannel: from humans to embodied agents. In *AISB*. University of Hertfordshire, Hatfield, UK, 2005.

[17] www.puredata.org

[18] The CMU Sphinx open source speech recognizer http://cmusphinx.sourceforge.net

[19] De Cheveigne, A., Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustic Society of the America* 111. 2002.

[20] P. Taylor. The Tilt Intonation model, *ICSLP 98*, Sydney, Australia. 1998.

[21] B.M. Streefkerk, L. C. W. Pols, L. ten Bosch, Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANNs, *Proc. Eurospeech'99*, Vol. 1, Budapest, 551-554, 1999.

[22] J.M.B. Terken, Fundamental frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America*, 95(6), 3662-3665, 1994.

[23] N. Obin, X. Rodet, A. Lacheret-Dujour, French prominence: a probabilistic framework, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP08)*, Las Vegas, U.S.A, 2008.

[24] V. K. R. Sridhar, A. Nenkova, S. Narayanan, D. Jurafsky, Detecting prominence in conversational speech: pitch accent, givenness and focus. In *Proceedings of Speech Prosody*, Campinas, Brazil. 380-388, 2008.

[25] D. Wang, S. Narayanan, An Acoustic Measure for Word Prominence in Spontaneous Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, Volume 15, Issue 2, 690-701, 2007.

[26] P. Viola, M.J. Jones, Robust Real-Time Face Detection, *International Journal of Computer Vision*, 137-154, 2004.

[27] C.G. Harris, M.J. Stephens, A combined corner and edge detector, *Proc. Fourth Alvey Vision Conf.*, Manchester, 147-151, 1988

[28] B. Lucas, T. Kanade, An Iterative Image Registration Technique with an Application to Stereo Vision, *Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 674-679, 1981.

[29] B. Baillie, URBI: Towards a Universal Robotic Low-Level Programming Language, *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems - IROS05*, 2005.

[30] D.M. Dehn, S. van Mulken, The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies*, 52: 1-22, 2000.

[31] Z. Ruttkay, C. Pelachaud, From Brows to Trust - Evaluating Embodied Conversational Agents, *Kluwer*, 2004.

[32] S. Buisine, J.-C. Martin, The effects of speech-gesture co-operation in animated agents' behaviour in multimedia presentations. *International Journal "Interacting with Computers: The interdisciplinary journal of Human-Computer Interaction"*. 19: 484-493, 2007.

[33] Dan R. Olsen, Michael A. Goodrich, Metrics for Evaluating Human-Robot Interactions. *Performance Metrics for Intelligent Systems Workshop held in Gaithersburg*, 2003.