SEMAINE

THE SENSITIVE AGENT PROJECT

D6c

Record interactions with automatic system

Date: 17 December 2010

Dissemination level: Confidential

ICT project contract no.	211486
Project title	**SEMAINE** **Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression**
Contractual date of delivery	*31 December 2010*
Actual date of delivery	*17 December 2010*
Deliverable number	D6c
Deliverable title	Record interactions with automatic system
Type	Report
Number of pages	13
WP contributing to the deliverable	WP 6
Responsible for task	Marc Schröder (marc.schroeder@dfki.de)
Author(s)	Gary McKeown, Roddy Cowie (QUB) Michel Valstar (ICSTM)
EC Project Officer	Philippe Gelin

Table of Contents

Executive Summary

The present report covers the recordings made of the dialogues between the fully automatic SAL system and a large number of human participants. The recordings of interactions with the fully automatic SAL system add substantially to the already available set of training data generated during previous phases of the project. One important addition of the new recordings is that all actions made by the automatic SAL system have been logged, which makes it possible to assess in detail how and why the system behaved the way it did. As the recordings are actual interactions with a conversational agent they serve very well as training material for future components.

All of the participants interacted with the system at two levels: a good version and a degraded minimal version. The participants have been asked to rate the quality of the interactions, and an observer rated the level of engagement live during the recordings. This provides the ability for strong statistical analysis of many aspects of human machine interaction. This analysis is provided in deliverable D6d.

In total, 81 participants have been recorded, for a total of 625 character interactions, totalling approximately 1860 minutes. Combined with the existing SEMAINE recordings these data constitute a considerable resource that should serve the affective computing and broader research community for many years to come.

Introduction

In year one and two of the SEMAINE project, we have created a large audiovisual database as part of an iterative approach to building agents that can engage a person in a sustained, emotionally coloured conversation, using the Sensitive Artificial Listener (SAL) paradigm. The data collected in the first two years was used to build the automatic SAL system. That data came from interactions between users and a human 'operator' simulating a SAL agent, and is described in detail in McKeown et al. 2010.

As part of an iterative design cycle, it is necessary to periodically evaluate the automatic SAL system, which will lead to modifications to the design. Because a large part of the automatic SAL system is data driven, this means that in order to create the modifications new data has to be generated. This data has to be suitable for the next development round, i.e. it should contain examples of conversations that the next iteration of the automatic SAL system is supposed to be able to have with a human user.

This report describes a new set of recordings done using the automatic SAL system. It consists of one prolonged set of experiments, that had both the goal to evaluate the current system and generate data to aid further development of the SEMAINE system. To allow both goals to be achieved simultaneously, the experiment used a within participants design in which each human participant interacted with two versions of the automatic SAL system. In the first interaction the participant talked to the latest version of the automatic SAL system, as developed by the SEMAINE team. In the second interaction the participants talked to an artificially degraded version of the SAL system. In each of these sessions they interacted with all four characters providing a total of eight character interactions for each participant. This way, it was possible to assess how well the automatic SAL system performed compared to a baseline. The recordings of both interactions but particularly the first set will also serve as data that can be used to develop the next generation of automatic SAL systems.

1 Fully automatic SAL system

In the fully automatic SAL recordings, the utterances and behavioural actions executed by the SAL Character were decided entirely automatically by the latest versions of the SEMAINE project system available. There have been three iterations of this procedure using different versions of the fully operational version of the SAL system. The details of which system was used are provided in the conducted experiments section.

2 Degraded SAL system

In a continuation of previous experimental methods used in the semi-automatic SAL scenario a comparison between full versions of the system and degraded versions of the system was considered the strongest approach to assessment. Two degraded versions of the system were used. The first degraded system used the random2face configuration file for the Java components and operated with the vision capabilities of the SAL system turned off. The effect of these was to provide a system that had random emotional prediction, that is it had unreliable information concerning the emotional state of the user.

The second degraded system was more extensively degraded with the goal of providing a system with as much of the affective computational capabilities removed as possible while still retaining

the key conversational elements of the SAL scenario. The SAL characters' voices were manipulated so that the system spoke with flat affect synthesised voices. The Greta facial expression and head action components were largely turned off, which means that no mimicking, no back-channeling behaviours occurred. Only animations of the voice (i.e. lip synching) and blinking remained. The dialogue management components were adjusted such that utterances were selected randomly and the decisions when to interject with an utterance would be made randomly based on a uniform random distribution over a certain interval.

3 Pilot Interactions

A number of interactions took place that were used as part of readying the system and preparation, these recordings were not formally part of any evaluation experiments but are made available as recordings as they usefully increase the volume of human-machine interactions and are suitable for training and evaluating many aspects of affective computing and/or human machine interaction in general and a number of aspects of the SEMAINE system in particular. There were more pilot interactions than were recorded as most of the piloting took place before the new recording set-up was fully operational. However 5 pilot interactions were actually recorded. In total 90 participants interacted with the system as part of the evaluation phase. 81 of these interactions were recorded and 76 of these were part of the formal experiments. The additional 5 pilot recordings took place between experiments 2 and 3 and were based on SEMAINE system 3.0.1 (revision 779).

4 Experiment Interactions

As with the original Solid SAL recordings, conducted in year one and two of the project, participants were informed of the goal of project, and explained how to use the system by the experimenter. An additional introduction was provided by the Greta character and is available with the recordings. The character interactions were limited to approximately 3 minutes. In total, five different settings were used in the experiment. An initial experiment (Experiment A) compared a version of the system based on SEMAINE system 3.0.1 (revision 734) with a degraded system in which visual feedback was turned off and user emotional state was randomly chosen (see Section 2). 16 participants were tested with this configuration adding 32 recording sessions to the database. A second experiment (Experiment B) used two different system versions; the full version was based on SEMAINE system 3.0.1 (revision 753) while the degraded system removed most of the affective cues from the system leaving only a stark basic SAL scenario with no back-channeling, emotional information and random utterance selection and flat affect in the agent voices. 30 participants were tested with this configuration adding a further 60 recording sessions to the database. A third experiment (Experiment C) used a different full version of the system based on SEMAINE system 3.0.1 (revision 782) this featured improved dialogue management and was compared with the same degraded system used in the second experiment. 31 participants were tested with this configuration adding a further 62 recording sessions. The full details of the number and type of recordings can be found in Table 1.

5 Recording setup

To allow for the recording of interactions with the automatic SAL, some modifications to the recording setup originally designed for the solid SAL recordings had to be made.

5.1 Sensors

Video is recorded at 49.979 frames per second and at a spatial resolution of 780 x 580 pixels using AVT Stingray cameras. Only the user is recorded in the automatic sessions from the front by both a greyscale camera and a colour camera. In addition the participant is recorded by a greyscale camera positioned on one side of the User to capture a profile view of their face and body. A greyscale AVT Stingray camera was also used for the video feature extraction components of the SEMAINE system.

The reason for using both a colour and a greyscale camera is directly related to the two target audiences. A colour camera needs to interpolate the information from four sensitive chip elements to generate a single colour pixel, while the greyscale camera needs only a single sensitive chip element. The greyscale camera will therefore generate a sharper image. Machine vision methods usually prefer a sharp greyscale image over a blurrier colour image. For humans however, it is more informative to use the colour image.

To record what the user is saying, we use two microphones: the first is placed on a table in front of the user and the second is worn on the head. The wearable microphones are AKG HC-577-L condenser microphones, while the room microphones are AKG C1000- S microphones. This results in a total of two microphones and thus two recorded channels. The wearable microphone is the main source for capturing the speech and other vocalisations made by the user, while the room microphones are used to model background noise. Audio is recorded at 48 kHz and 24 bits per sample.

5.2 Environment

The user is located in a separate room from the automatic system setup. The user can hear the system output (i.e. the SAL character's voice) over a set of speakers. They can see the Greta avatar through a screen reflected in a teleprompter. Within the teleprompter, the cameras recording a user's frontal view are placed behind the semi-reflecting mirror. This way, the user can have the sensation that they look directly at the SAL agent while still being recorded as if they were making eye-contact with the camera. Professional lighting is used to ensure an even illumination of the faces. The greyscale AVT Stingray camera used by the video feature extraction components of the SEMAINE system was also placed behind the semi-reflecting mirror.

5.3 Synchronisation

In order to do multi-sensory fusion analysis of the recordings, it is extremely important to make sure that all sensor data is recorded with the maximum synchronisation possible. To do so, we used a system developed by Lichtenauer et al. (2010). This system uses the trigger of a single camera to accurately control when all cameras capture a frame. This ensures all cameras record every frame at almost exactly the same time. The same trigger was presented to the audio board and recorded as an audio signal together with the four microphone signals. This allowed us to synchronise the audio and the video sensor data with a maximum time difference between data samples of 25 microseconds.

5.4 Data compression

The amount of raw data generated by the visual sensors is very high: 625 character interactions, lasting on average 3 minutes, recorded at 49.979 frames/second at a spatial resolution of 780*580 pixels with 8 bits per pixel for 3 cameras, would result in 6.9 TeraByte. This is impractical to deal

with: it would be too costly to store and it would take too long to download over the Internet. Therefore, the data has been compressed using the H.264 codec and stored in an avi container. The video was compressed to 440 kbit/s for the greyscale video and to 500 kbit/s for the colour video. The recorded audio was stored without compression, because the total size of the audio signal was small enough.

5.5 Agent recordings

Recordings of the agent avatar were made using the Fraps real-time video capture software (http://www.fraps.com/). These are similarly compressed using the H.264 codec and stored in an avi container. These are provided separately from the user recordings and are unfortunately not synchronised with the user videos.

6 Participants

In total 90 participants were tested . 14 of these were in piloting and preparation stages, 9 before experiment 1 and 5 between experiments 2 and 3. 16 participants took part in Experiment 1, 11 female and 5 male. 30 participants took part in Experiment 2, 19 female and 11 male. 31 participants took part in Experiment 3, 25 female and 6 males, one female withdrew due to ill health.

7 Stages in recordings

Each of the automatic system recordings is divided into a number of stages that represent the different parts of the experimental process. The first stage is the introduction in which the Greta character delivers a monologue introduction to the system and the characters, this part of the interaction is made available and could serve as a baseline or neutral expressionless face for many of the participants. Following this comes the first character interaction, a character is chosen at random by the system and the participant interacts with the character for approximately three minutes. The end part of the character interaction is chosen by an experimenter at the first occasion after three minutes that causes minimal disruption, however the termination of the conversations is often abrupt. The next phase is the evaluation question phase where the Greta character returns and offers the participant an opportunity to comment on the interaction. This is followed by the three evaluation questions. Once the evaluation phase is over the next character interaction begins and the process is repeated until the last character evaluation has taken place. The recording of the interaction then ends.

8 Annotations

As the evaluation of the final versions of the system necessarily occurs towards the end of the project the time constraints limit the amount of annotation that can be achieved. However a number of variables were recorded as part of the experimental process of the experiments and will be made available. Additionally, as the main form of annotation has involved continuous trace style ratings of interactions it was considered valuable to have at least some minimal trace ratings of the automatic interactions. These tracings occurred live as an expert tracer watched a screen displaying the human participant in a separate room. The rating that was chosen was level of engagement on a scale of 0 to 1. The text anchoring each end of the trace dimension was “Absolutely no sense of engagement” and “Compelling sense of engagement” with two additional textual markers “weakly engaged” placed a third of the way along the trace dimension and “quite engaged” at the two thirds

mark. Traces always started at the “quite engaged' point. Each interaction with a character was traced typically providing eight traces for each human participant and a total of 702 traces of character interactions.

Table 1 Recording Session Information (time is measured in minutes)

Experiment & System	**Sessions**	**Users**	**Approximate Total Character Interaction Time**	**Annotators**
Experiment A				
Full 1	60	15	180	1
Degraded 1	60	15	180	1
Experiment B				
Full 2	120	30	360	1
Degraded 2	120	30	360	1
Experiment C				
Full 3	121	31	360	1
Degraded 2	124	31	360	1
Pilots	20	5	60	1
Total	625	81	1860	1

9 Experimental Variables

As part of the experimental process a number of self report questions were asked at the end of the interaction to assess the quality of the interaction from the point of view of the participant. These variables will be made available with the database. The questions asked at the end of the character interaction were:

1. Could you state the way you felt about the conversation? You don't have to make any comments but feel free to comment if you want to.
2. First of all, how naturally do you feel the conversation flowed? Zero means not at all, ten, means totally natural.
3. Did you feel the Avatar said things completely out of place? If yes how often? Never, a few times, quite often, most of the time, all the time. (Scored, 5, 4, 3, 2, 1 respectively)
4. How much did you feel you were involved in the conversation? Zero means not at all, ten, means completely involved.

The first question resulted in open ended comments that can be informative of the way the participant felt about the conversation, not every participant commented on the sessions. Questions 2, 3 and 4 resulted in quantitative scores where a higher score means a better interaction. Questions 2 and 4 were scored from 0 to 10 and question 3 was scored from 1-5. Additionally at the end of both interactions participants were asked which system they preferred the full system or the

degraded system. There was also an overall rating of how good the session was by one external rater.

10 Automatic Analysis of the Recordings and Log Files

A number of variables are automatically extracted from the system during these interactions to serve as variables that can be used for analysis. The synchronous high quality audio and video streams allow the automatic analysis by the SEMAINE audio and vision systems. Jiang, Valstar and Pantic (in print), Gunes and Pantic (2010), Nicolaou and Pantic (2010) Eyben et al. (in print) report on some of the methods used in the automatic analysis of the data and how these automatic annotations were obtained. These data are of course limited by the abilities of the methods used to obtain them and any use should adequately account for this. These logs have been divided into character sessions for each of the experiments to be made available on the database. The variables are categorised depending on which part of the interaction they are relevant to. The user related variables included in the log files are shown in Table 2. The agent related variables included in the log files are shown in Table 3. The system related variables included in the log files are shown in Table 4. Not all the variables are implemented for each experiment.

Table 2 User related log file variables

time	Time in milliseconds since the the system started
u.headGesture	User head nods, tilts, shakes
u.headGestureStarted	Time of user head gesture start time
u.headGestureStopped	Time of user head gesture stop time
u.facialExpression	Higher level expression (not implemented)
u.facialActionUnits	FACS Action Units detected
u.facialExpressionStarted	Higher level expression start time
u.facialExpressionStopped	Higher level expression stop time
u.facialActionUnitsStarted	FACS Action Units start time
u.facialActionUnitsStopped	FACS Action Units stop time
u.pitchDirection	Pitch rises and falls
u.speaking	Boolean user speaking variable
u.vocalization	Breath, Laughter, Sigh detection
u.facePresent	Is a face detected or not?
u.interest	System estimation of the user's level of interest
u.valence	System estimation of the user's level of valence
u.arousal	System estimation of the user's level of arousal
u.potency	System estimation of the user's level of potency
u.unpredictability	System estimation of the user's level of unpredictability
u.intensity	System estimation of the user's level of intensity

u.emotion-quadrant	System estimation of the user's emotional quadrant (e.g. Positive Active)
u.userUtterance	System estimation of the user utterance
u.userUtteranceStartTime	user utterance start time
u.userUtteranceFeatures	user utterance features (e.g. short, normal, agreement)
u.gender	System estimation of the user's gender
u.buttonEvent	Yuck button event
u.buttonEventTime	Yuck button event time

Table 3 Agent related log file variables

a.needToSpeak	Agent's need to speak
a.turnTakingIntention	e.g. stopSpeaking startSpeaking
a.agentUtterance	Text of agent utterance
a.agentUtteranceStartTime	Agent utterance start time
a.agentHead	Agent head gesture e.g. head_up_left, head_nod
a.agentHeadStartTime	Agent head gesture start time
a.agentFace	Agent face gesture e.g. close-eyes, raise_eyebrows
a.agentFaceStartTime	Agent face gesture start time
a.agentGaze	Agent gaze direction e.g. look_down_left, look_at
a.agentGazeStartTime	Agent gaze direction start time
a.agentGesture	Agent gesture
a.agentGestureStartTime	Agent gesture start time
a.agentTorso	Agent torso
a.agentTorsoStartTime	Agent torso start time
a.agreement	Agent agreement level (set for each character)
a.acceptance	Agent acceptance level (set for each character)
a.belief	Agent belief level (set for each character)
a.liking	Agent liking level (set for each character)
a.understanding	Agent understanding level (set for each character)
a.interest	Agent interest level (set for each character)
a.anger	Agent anger level
a.sadness	Agent sadness level
a.amusement	Agent amusement level
a.happiness	Agent happiness level

a.contempt	Agent contempt level
a.anticipation	Agent anticipation level
a.solidarity	Agent solidarity level
a.antagonism	Agent antagonism level

Table 4 System related log file variables

d.userTurnState	User turn state
d.agentTurnState	Agent turn state e.g. listening, speaking, expectingAnswer
d.convState	Conversation State
c.userPresent	Is the user present?
c.character	Current Character
c.nextCharacter	Next Character
c.dialogContext	e.g. AnnounceNextCharacter, Questions1

11 Overview

The recordings of interactions with the fully automatic SAL system add substantially to the already available set of recordings from previous phases of the project. As they are actual interactions with a conversational agent they serve very well as training material for future components. They provide a range of levels of interaction, with participants reporting very enjoyable and positive interactions with the system in some of them and there not being any interaction at all between the participant and the agent in others. A substantial sample interacted with varying versions of the system. All of the participants interacted with the system at two levels: a good version and a degraded minimal version providing the ability for strong statistical analysis of many aspects of human machine interaction. Combined with the existing SEMAINE recordings these constitute a considerable resource that should serve the affective computing and broader research community for many years to come.

12 References

Eyben, F., Wöllmer, M., Valstar, M.F., Gunes, H., Schuller, B. and Pantic, M. “String-based audiovisual fusion of behavioural events for the assessment of dimensional affect,” in Proc. IEEE Int’l conf. Face and Gesture Recognition, 2011, accepted for publication.

Gunes H. and Pantic, M. (2010) “Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners,” in Proc. of the International Conference on Intelligent Virtual Agents, pp. 371–377.

Jiang, B., Valstar, M. and Pantic, M. “Action unit detection using sparse appearance descriptors in space-time video volumes,” in Proc. IEEE Int’l conf. Face and Gesture Recognition, 2011, accepted for publication.

Lichtenauer, J., Shen, J., Valstar, M. and Pantic, M. (2010) “Cost-effective solution to synchronised audio-visual data capture using multiple sensors,” in Proc. IEEE Int’l Conf’ Advanced Video and Signal Based Surveillance, Nov 2010, pp. 324–329.

Gary McKeown, Michel F. Valstar, Roderick Cowie, and Maja Pantic, 'The SEMAINE Corpus of Emotionally Coloured Character Interactions', Proc. IEEE Int’l Conf. Multimedia & Expo (ICME’10), pp. 1079-1084, Singapore, July 2010

Nicolaou, H. G. M. A. and Pantic, M. (2010) “Automatic segmentation of spontaneous data using dimensional labels from multiple coders,” in Proc. of LREC Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, 2010, pp. 43–48.