

Multimodal Corpora for an Automatic System Fostering Participants' Engagement in Informal Conversations around a Museum Café Table

Nadia Mana, Alessandro Cappelletti, Oliviero Stock and Massimo Zancanaro

FBK-irst

via Sommarive 18

38123 Trento, Italy

{mana,cappelle,stock,zancanaro}@fbk.eu

ABSTRACT

In this paper we present the multimodal data collected for developing a system able to influence the behavior of small groups in an informal and non goal-oriented conversation scenario. The prototype system looks like a table in a museum cafeteria and it is aimed at inducing the people sitting around to talk about their visit to the museum. To this aim, the system provides visual cues to foster participants' engagement in the conversation. The cues are contextualized by automatically monitoring the group dynamics and by continuously planning and executing minimalist strategies based on the participants' speaking activity and visual attention. In the paper, we shortly describe the system, its main components and functionalities. We then present the two data collections carried out to gather multimodal data to tune the basic perceptual modules of the system (voice activity detector and face tracker) and to improve the presentation engine of the visual cues.

Keywords

Script-based behavior, spontaneous conversation, behavior monitoring, speech activity, head orientation, visual attention.

1. INTRODUCTION

In this paper, we present the multimodal data collected to develop and test a system that aims at influencing the behavior of the participants in an informal, non goal-oriented, co-located small group conversation.

Specifically, the system is meant to be located in a museum café, seen as the place where people take a rest after the visit and as a natural setting where to enhance the museum experience of the group.

The choice of targeting a museum scenario is driven by a growing evidence of the fact that the social dimension of the museum visits (museum visitors are rarely isolated individuals and more often part of a group), strongly affects the museum experience and it plays a major role in successful learning engagement. Such engagement may depend on how much time, effort and collaborative attitude are spent in conversing and commenting on what has been seen [9]. As discussed in [6], computers may become persuasive tools able to influence people's behavior and our ultimate objective is to experiment with a persuasive system in the context of informal learning. Since our system is supposedly not invasive, we designed it as a tabletop display that can accommodate a small group of people after the museum visit. The system continuously monitors the group dynamics and uses the accumulated knowledge to plan and deploy minimalist

strategies to influence behavior through indirect means. The system is not interactive in a common sense. In this way, the main "interaction channel" is left for direct human-to-human interaction and neither major conscious elaboration effort nor actions by the user towards the interface are required.

Similarly to peripheral displays [15], our system interface is not central to the attention of the group and people may look at it only occasionally. However, while the peripheral displays have mainly a passive role (usually aiming at providing easily graspable information to the users), our system is meant to stay at the periphery most of the time and appropriately attracts the users' attention, according to the context and the goals of engagement fostering.

The usage of minimalist strategies in presenting information, in turn, takes inspiration from the world of advertising with its resort to evocative, mainly visual messages to influence behavior ([11],[5]).

More in general, according to [12], our work is inspired to the elaboration likelihood model (ELM) of persuasion, based on two ways of changing people's attitudes: the central route and the peripheral one [16]. According to ELM, the central route has a person attentively attending to an argumentative communication, and pondering about the ideas and concepts presented, while the peripheral route influences people by means of peripheral cues (e.g., the status of the communication source, its credibility, its attractiveness, etc.). The relative effectiveness of the two routes depends on (a) the listener's motivation to pay attention and think and (b) the time on the listener's hands. The central route requires the target person(s) to be ready to attend an argumentative communication and to have time to go through possibly complex argumentation chains. On the contrary, the peripheral route is more amenable to situations in which people are involved in different matters and/or whenever it is not appropriate to distract them from their current focus. Moreover, that route can unfold its effects in much shorter time.

In order to test and tune the basic perceptual modules of the system and to improve the system performance, we carried out two multimodal data collections: the first one to collect script-based examples of speech activity and visual attention; the second one to collect examples of spontaneous conversations with the system prototype working.

In the next sections, after a presentation of related works we briefly describe the system while presenting its perceptual components, presentation module and graphical interface (see [16] for more details about the system). Then we present the two data collections, focusing on the collected data, the experimental

procedure and the annotations done. Finally, we shortly discuss the quality of the collected data, also in relation to limitations due to the basic technology and the applicative scenario.

2. RELATED WORK

There have been several studies on the display of information through social tools. For example, Groupcast is a wall projected office application that creates informal interaction opportunities by displaying mutual interest to people passing by [10]. Drift is an interactive table that displays an aerial photo of England through a hole, to foster interpretation and engagement [7]. Qualitative observations showed that people got engaged by interacting with the system and narrating about the places spotted. Hello Wall is a digital wall made of a grid of lights [13]: depending on people distance, the wall changes communicative function (ambient, notification, interaction). Abstract light patterns convey information about mood, presence and crowdedness.

DiMicco and Bender [4] have experimented with a system that, by monitoring working groups, presents information about relational behavior in the form of graphical displays on a tabletop device, to affect group behavior.

A similar approach was pursued by Sturm and colleagues [14] who used a tabletop device as a peripheral display aiming at the same self-regulatory effect as discussed above. In their approach, they display not only the speaking time but also the gaze behavior of their participants. Their results show a similar effect of DiMicco and Bender [4] for what concern the speaking behavior and no effect on gaze behavior.

Kim and colleagues [8] used a portable device called ‘the sociometric badge’ to monitor speaking activity and other social signals in a team. They also used portable display to provide an individual graphical representation of the group behavior. Their results showed a reduction of the overlapping speech but not a significant increase in solo speech condition.

All these approaches are based on the idea that reflection on one’s own behavior may bring to rational decisions about behavior changing [2]. These systems are usually applied in a team-work scenario where each participant is motivated to achieve his/her goal (e.g., a successful meeting and/or a well-accepted personal appearance).

Our work is similar to some extent because of the use of a display to provide contextually appropriate information but it is slightly different because we aim at providing a calm visualization of stimuli aimed at fostering the conversation of the group rather than balancing the participation.

3. SYSTEM DESCRIPTION

The system is built as a tabletop display (see Figure 1), where the group dynamics is continuously monitored by a speaking activity tracker and a visual attention detector based on a face tracker.

The surface of the table is used as a projection display for calm visualizations so as not to attract the users’ attention, unless it is explicitly intended by the system to do so: a pond is displayed in the center of the table with red fishes quietly swimming around (see Figure 2).

The stimuli presented are pictures of exhibits and sheets of paper with short text, all floating on the water. An increase of activity of the fish swimming close to a given stimulus, or water drops

falling close by are used to attract the participants’ attention toward it.



Figure 1. Participants while using the system

The system working is based on two kinds of information: each participant’s attention, as estimated by a face detector processing the video streams from four cameras; and the participants contribution to the conversation estimated by a Voice Activity Detector (VAD) using the audio from four microphones positioned on the table.



Figure 2. The pond displayed on the graphical interface of the system

3.1 Architecture

The system is composed of two main parts: the perceptual modules and the presentation engine.

The perceptual modules process data from the visual and the acoustical scene as provided by 4 cameras hidden at the center of the table, and 4 microphones placed in front of each participants. In particular, the speaking activity of each participant is analyzed and tracked according to the output of a VAD and a label (‘Speaking’/‘NotSpeaking’) is assigned. At the same time, on the visual attention side, a commercial face tracker is used to track the head orientation, characterized by specific values of 3D coordinates, pan, tilt and rotation, recorded every second (on average).

Presentation generation is based on the system’s understanding of the group behavior and is driven by appropriate communicative strategies. It is a rule-based engine where the rules can be

conditioned using the variables that represents the group's and the individuals' conversational and attentional states.

A rule is a sequence of presentational instructions with a precondition as a logical expression on the contextual variables. The engine is based on a 3-step process (where the next cycle is performed as soon as the triggered rule is finished) in which (i) the sub-set of rules with the precondition at true are activated; (ii) a single rule is chosen with a tie-break decision, and (iii) the rule is triggered.

In the present version, four rules have been implemented:

- R1: if somebody looks at the table and does not talk, while somebody else is talking: present to the person, who is neither speaking and nor looking at the table, his/her own topic (i.e. a specific topic related to the museum visit, virtually assigned to him/her by the system at the session beginning).
- R2: if somebody looks at the table and does not talk, while somebody else is talking: present the topic of the person who is talking to the person who is not talking but is looking
- R3: if nobody is talking and nobody is looking at the table: present to the person who is speaking the least his/her own topic
- R4: if nobody is looking at the table, somebody is talking and somebody else is not talking: present the topic of the person who is speaking to the person who is neither speaking and nor looking at the table

Each rule introduces one or more visual stimuli and addresses them to one or more individuals (especially who is engaged the least in the conversation in order to capture their attention, to stimulate their talk and to foster their participation). Upon completion, the stimuli are removed from the table so that each rule is independent from the others. This is of course a limitation for the implementation of complex presentation strategies that span multiple rules.

3.2 Graphical Interface

As said, the graphical interface of the system represents a pond, where floating pictures of exhibits and sheets of paper with short text are the stimuli presented to the people at the Augmented Café Table (see Figure 3).



Figure 3. Example of stimuli presented

At present, the table has a database of 4 topics related to the frescoes of Torre Aquila, a medieval frescoed room in Buonconsiglio castle in Trento). For each topic, the system can access 14 images (the whole frescoes and 13 details), each one described by 4 textual descriptors (keywords or short sentences). When a topic is selected by the system, the latter randomly retrieves one of the images and two textual descriptors related to that topic and displays them on the table with an animation. The system takes care of avoiding repetition of rules and topics in subsequent rules and randomizes images and descriptors.

4. DATA COLLECTION

While aiming at developing the system described in Section 3, two data collections were carried out in a laboratory in order to gather multimodal data to: a) tune the basic perceptual modules of the system (voice activity detector and face tracker); and b) to improve the presentation engine of the visual cues.

The first one was script-based, whereas the second one aimed at collecting examples of real use of the system prototype with spontaneous conversations.

The script-based data collection involved 10 groups of 4 people (24 male and 21 female), whereas the second data collection involved 6 groups of 4 people (16 female and 8 male). All participants were recruited among the FBK employees. Most of them came from the administration. We grouped the participants making attention to their professional roles in order to avoid hierarchical influences on the conversational behavior. Furthermore, to make the experimental scenario more similar to the real one (where people visiting a museum in group are more or less familiar with each other), no groups were composed by people completely unknown. Still, similarly to a real scenario, the participants' familiarity with the Torre Aquila frescoes was quite varied. Finally, none of the people recruited for the second data collection participated to the previous one.

4.1 Script-based Experiments

The participants (4 per session) were asked to sit in pre-defined positions around the table and instructed to enact some specific behaviors at a given time. In particular, the experiment was composed by three tasks.



Figure 4. A session of the script-based experiments (Task 1): head orientation toward the right

In the first task, the experimenter asked to participants to look at a specific point on the table (a small colored square placed close to the subject or those close to the opposite participant) or a

specific person (e.g., the person in the front, that on own left, or that on own right – as depicted in Figure 4) for 5 seconds.

In the second task each participant was asked to individually introduce him/herself by saying own name, where he/she lived and his/her preferred color. During this task, the speaking participant was looking at the opposite subject, while the other participants, silently listening the self-presentation, were looking at the speaker.

Finally, the third task consisted of two parallel sub-tasks (schematized in Figure 5): participant 2 was talking to participant 1, while simultaneously participant 3 was talking to participant 4. During the task, participant 1 and 2 were frontally looking at each other, whereas participant 3 and 4 were looking at a green square placed on the table.

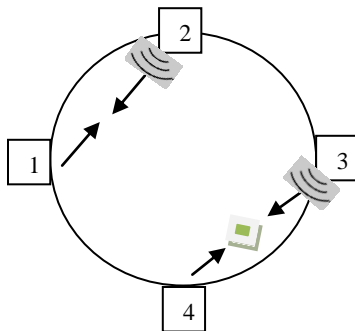


Figure 5. Scheme of Task 3: parallel speaking

During all these experiments the presentation component of the system was disabled and just the perceptual modules were running.

4.2 Experiments based on spontaneous conversations

In the experiments of the second data collection the table surface was covered to hide microphones cables and the four cameras at the center of the table were hidden to make the technology less invasive and the Augmented Café Table more similar to a common table (see Figure 6).

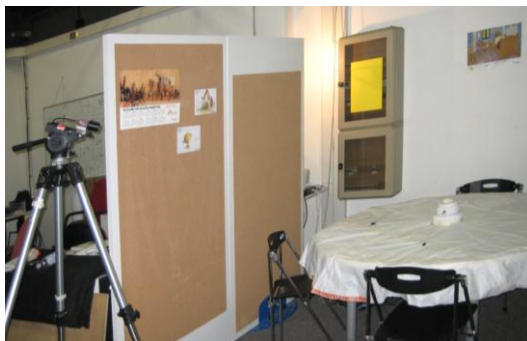


Figure 6. Technical set-up of the second data collection

This data collection was designed as a within-group study with two conditions: the “Contextual” condition used the contextual information provided by the continuous automatic behavioral monitoring to trigger the rules (as described in Section 3.1); the other version (called “Random”), used as control condition, was based on the same presentation rules but triggered on a random

basis: the frequency of the activation of the rules was set in such a way to generate in average the same rules triggered in the contextual condition.

At the beginning of the experiment, the participants were invited to visit the reproduction of four frescos of a local castle (see Figure 7). They were told that the objective of the study was to investigate the effectiveness of different ways of describing the frescos. To help them in this task, each one received a short description of two of the frescoes (but partially different from the description of the other participants to stimulate the discussion).



Figure 7. A participant during the first phase of the second data collection

After this introductory phase, the four participants were invited in the Augmented Café Table room for a debriefing interview, with the excuse of the lack of meeting rooms available. Then, the experimenter existed the room with a pretext and left the group for about 20 minutes (time sufficient to have a natural and lively conversations, estimated by preliminary experiments). During this time, the two versions of the system were run in sequence. The order was randomized: half of the groups experimented first the contextual condition followed by the random condition and half of the groups the other way round.

All the experimental sessions were video recorded by an external camera, laterally placed (see Figure 6).

5. AUTOMATIC AND SEMI-AUTOMATIC ANNOTATIONS

During the script-based data collection, the collected examples were annotated by means of an annotation console (see Figure 8), where the experimenter was marking start and end time of each specific exercise asked to the participants by specific buttons.

The annotation produced an XML file per each experimental session, where the temporal information of each task (Task 1, Task 2 and Task 3) and sub-task were automatically saved (see Figure 9).

On the acoustic and visual sides, in both the data collections, according to the output of the perceptual modules, the speaking activity and visual attention were automatically annotated (on average every second) per each participant, using the Speaking/NotSpeaking and Front/Left/Right/onTable labels respectively.

In the second data collection, an automatic annotation was also added. It was concerning the stimuli presentation and their

activation rules: which rule was fired, its activation duration, the topic of the presented stimuli and which participant it was targeting. These data, put in relation with the output of the perceptual modules, will make other automatic analyses of the conversations and the participants' behaviors possible (e.g., investigating participants' speech activity and visual attention in correspondence of stimuli presentation).

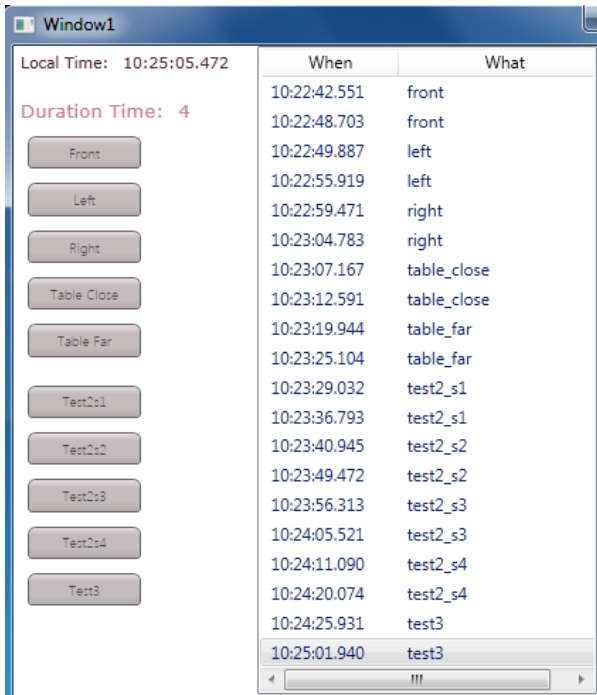


Figure 8. Interface of the annotation console used during the script-based experiments

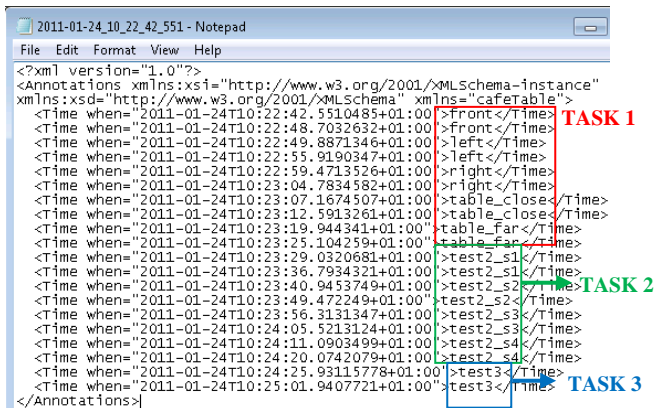


Figure 9. Example of XML annotation of an experimental session script-based

6. DISCUSSION

We avoid here any detail concerning the basic perceptual modules and system assessment (because it is not the major goal of this paper), but report some considerations about the collected data and the problems surfaced during the two data collections.

Although the script-based examples are limited because of the shortness of the each task (5-10 seconds) and the lack of spontaneity in the participants' behavior, such data are very useful to assess and tune the basic perceptual modules (VAD and face tracker), as well as to train machine learning algorithms to improve the performance of the visual attention detector. In particular, given the temporal information extracted from the XML annotation, we can automatically select the data related to specific tasks (e.g., all the examples of head oriented toward right, or all the example of simultaneous speaking, etc.) from the automatic annotations based on the perceptual modules output.

On the other side, the examples of conversational behavior while the system was working offer the advantage of being longer and spontaneous. However, even if both the data collections were carried out in controlled conditions (stable lighting and reduced environmental noise), in the second one the performance of the basic components suffered of some limitations, indeed due to the experiment naturalness and spontaneity. In particular, on the acoustic side, the voice of a specific person (especially if she/he had a high tone of voice) was very often also recorded by the microphone of the adjacent participants. Furthermore, in many cases during the experiment people sprawled on the chair and the microphone positioned on the table could not capture their voice. These problems could be solved by using close-talk microphones but we avoided them on purpose because we are interested in investigating the possibility of developing systems that are (fully or at least partially) not invasive (also in prospect of a real museum setting).

Also on the visual side we met some problems: sometimes a subject was out of reach of the camera because of the posture. This situation usually went on for few seconds and just in few cases persisted for several minutes, severely impacting on the quality of the collected data (besides the average accuracy of the system). More in general, the most problematic situation was the inaccurate positioning of the mask on the subject face (see Figure 10): in these cases, the system often persisted in tracking the person in a wrong way until the person was lost by the camera and then tracked again. In other cases, the problem was due to the difficulty of the tracker in detecting the subject's face because of occlusions (e.g., hands on the face). Again, this problem usually tended to persist until the subject changed position removing any occlusion. In both situations, the collected data were inevitably incorrect.

Since the specific face tracker we used in the system does not allow to save the video but only the tracking, we cannot consider the possibility of correcting the annotation in a post-processing phase. Alternatively, other face or eyes trackers (e.g., [1] and [3]) could be used.

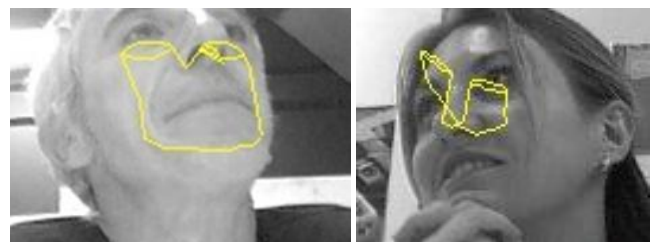


Figure 10. Examples of misaligned and incorrect mask

The problems on the acoustic and visual tracking side mentioned above were definitely less evident in the first data collection, indeed for the more controlled situation due to the script-based participants' behavior.

In conclusion, the script-based data are more complete and correct but the data of the experiments based on spontaneous conversations are more similar to the real application scenario.

7. CONCLUSION

In this paper we have presented two data collections carried out within the research activity that aims at developing a system able to influence the behavior of a small group of people involved in an informal and non goal-oriented conversation. The participants' behavior is meant to be influenced by providing them visual cues to foster the engagement in the conversation.

The system, meant as an Augmented Café Table, is composed of two main parts: a) the perceptual modules, processing data provided by microphones and cameras, and b) a presentation engine, based on the system's understanding of the group behavior and driven by appropriate communicative strategies, that generates animated visual stimuli, projected on the table.

The goal of the first data collection was to collect script-based examples of speech activity and visual attention, whereas the goal of the second one was to collect examples of spontaneous conversations and behaviors while the system prototype was running.

The collected data are useful to test and calibrate the basic perceptual modules of the system, namely the speaking activity tracker and the visual attention detector (based on a VAD and a face tracker, respectively), as well as to tune the system and to improve its performance, especially on the presentation and rule activation side. Furthermore, the data can also be used to train possible machine learning algorithms (e.g., to develop an automatic engagement detector), to be included in the system.

In addition, the audio-video recordings can be used for other research purposes (e.g., spontaneous multi-party speech recognition, conversation analyses, focus of attention, etc.).

Finally, we are considering the possibility of extending our corpora by taking the system into a museum cafeteria and collecting data in a real setting.

The annotation dataset of the two data collections and the audio (.wav files) and video (.avi files) recordings of the spontaneous conversations of the second data collection are freely available for the research community on demand.

8. REFERENCES

[1] Ba S.O. and Odobez J.M. Recognizing Human Visual Focus of Attention from Head Pose in Meetings, *IEEE Trans. on Systems, Man, and Cybernetics*, April 2008.

[2] Boud D., Keogh R., and Walker D. (Eds.). *Reflection: Turning Experience into Learning*. Kogan Page (1985).

[3] Brutti A. and Lanz O. A joint particle filter to track the position and head orientation of people using audio visual cues. *European Signal Processing Conference - EUSIPCO 2010*, , Aalborg, Denmark, pp. 974-978 (2010).

[4] DiMicco J.M., and Bender W. Group Reactions to Visual Feedback Tools. *PERSUASIVE 2007, LNCS 4744*, pp. 132-143 (2007).

[5] East R. *The Effect of Advertising and Display: Assessing the Evidence*. Boston: Kluwer Academic Publishers (2003).

[6] Fogg B. J., *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan-Kaufmann (2002).

[7] Gaver W. W., Bowers J., Boucher A., Gellerson H., Pennington S., Schmidt A., Steed A., Villars N., and Walker B. The drift table: designing for ludic engagement. *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, pp. 885-900 (2004).

[8] Kim T.J., Chang A., Holland L., and Pentland A.S. Meeting Mediator: Enhancing Group Collaboration with Sociometric Feedback. *CHI '08 extended abstracts on Human factors in computing systems*, pp. 3183-3188 (2008).

[9] Leinhardt G., and Knutson K. *Listening in on Museum Conversations*. Altamira Press, (2004).

[10] McCarthy J.F., Costa T.J., and Liongosari E.S. UniCast, OutCast & GroupCast: Three Steps Toward Ubiquitous, Peripheral Displays. *Proceedings of the 3rd international conference on Ubiquitous Computing*, pp. 332-345 (2001).

[11] Messaris, P. *Visual Persuasion: The Role of Images in Advertising*. Thousand Oaks, California: Sage Publications (1996).

[12] Petty R.E., and Cacioppo J.T. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer-Verlag (1986).

[13] Streitz N.A., Rucker C., Prante T., Alphen D.V., Stenzel R., and Magerkurth C. Designing Smart Artifacts for Smart Environments. *Computer*, vol. 38, no. 3, pp. 41-49 (2005).

[14] Sturm J., Houben-van Herwijnen O. Eyck A., and Terken J.. Influencing Social Dynamics in Meetings through a Peripheral Display. *Proceedings of the 9th International Conference on Multimodal Interfaces*, pp. 263-270 (2007).

[15] Weiser M., and Brown J.S.. *Designing Calm Technology*. *PowerGrid Journal*, vol. 1, pp. 1-17 (1996).

[16] Zancanaro M., Stock, O., Tomasini D., and Pianesi F.. A Socially Aware Persuasive System for Supporting Conversations at the Museum Café. *Proceedings of the 16th international conference on Intelligent user interfaces*, pp. 395-398 (2011).