MAHNOB HMI IBUG MIMICRY DATABASE (MHI-MIMICRY)

JEROEN LICHTENAUER, MICHEL VALSTAR, XIAOFAN SUN, ANTON NIJHOLT MAJA PANTIC

1. Introduction

The experiments from which this database contains the recordings, were conducted with the aim of the analysis of human interaction, in particular mimicry, and elaborate on the theoretical hypotheses of the relationship between the occurrence of mimicry and human affect. The recorded experiments are designed to explore this relationship. The corpus is recorded with 18 synchronised audio and video sensors, and is annotated for many different phenomena, including dialogue acts, turn-taking, affect, head gestures, hand gestures, body movement and facial expression. Recordings were made of two experiments: a discussion on a political topic, and a role-playing game. 40 participants were recruited, all of whom self-reported their felt experiences. The corpus will be made available to the scientific community.

1.1. Experiment Aim.

As a general starting point in our database design, specific hypotheses in the field of social psychology determined what kind of scenarios would be suitable for our recordings. We chose to design our recording scenario such that the collected data allows us to test two hypotheses about mimicry that have been posed in the literature:

Hypothesis 1. Agreement-/Disagreement-Mimicry occurs in conversations when the participants agree with each other as well as when they do not agree with each other, with a higher frequency or amount of mirroring during agreement than during disagreement. Moreover, mimicry occurs in conversations in with there is the intention to gain acceptance from an interaction partner through conforming to that persons attitudes, opinions, and behaviours.

Hypothesis 2. Affiliation-Mimicry has the power to improve social interaction. That is: when individuals communicate, one partner who wants to affiliate with others may intentionally engage in more mirroring of them; in contrast, when they want to disaffiliate they intentionally engage in less mirroring.

Based on the theoretical foundations of the above two mimicry hypotheses, we designed two conversational scenarios. The first scenario is a debate, and the second scenario is a role-playing game where one participant plays the role of a homeowner who wants to rent out a room, and the other participant plays the role of a student who is interested in renting the room.

For more background on the motivations for these experiments, as well as references to related literature, please refer to [2].

London, 2011.

1.2. **Experiment Form.** The recording includes two experiments. In Experiment A, participants were asked to choose a topic from a list. Participants were then asked to write down whether they agree or disagree with each statement of their chosen topic. The discussion is held between the participant and a confederate. Participants are led to believe that the confederate is a fellow nave participant. Participants were asked to start the conversation by presenting their own stance on the topic, and then to discuss the topic with the other person, who may have different views about the topic.

Every topic has a list of statements regarding that topic associated with it. In the pre-recording assessment, the participants note their (dis)agreement with these statements. This is used as a reference for annotating, possibly masked, opinion or attitude. During the discussion participants and confederates express agreement and disagreement, and show a desire to convince the other person of their opinion.

In Experiment B, the intent was to simulate a situation where two participants want to get to know each other a bit better and need to disclose personal and possibly sensitive information about them in the process. Participants were given a communication assignment that requires self-disclosure and emotional discovery. Participant 1 played a role as a student in university who was looking for a room to rent urgently Participant 2 played a role as a person who owns an apartment and wants to rent one of the rooms to the other one.

Participants are not sure about their partners preference at the beginning, so the hypothesis is that they will try to get more information from their partners first, only gradually showing more sensitive personal information to the other. Moreover, their conversation partners may not want to expose many details to them until s/he decides whether the participant is someone they like or not. However, they have the same goal, which is to share an apartment, so they have the tendency of affiliation.

To rule out mixed gender effects, experiments included either all male participants and confederates, or all female. After recording both sessions, participants finished a personality questionnaire and two separate experiment questionnaires, which were designed to measure the experienced affect and attitude during the two sessions.

1.3. **Self-Report of Participants.** 28 male and 12 female students from Imperial College London (aged 18 to 40years) are participants. Each was paid 10 pounds for participating in the study, which took about 1.5 hours. Two male confederates and one female confederate were from the iBUG group at Imperial College London. All participants were assigned to each other randomly. Four personality questionnaires were finished before attending the experiment. These were: 1) Big-Five Mini-Markers FNRS 0.2, 2) The Aggression Questionnaire including four subscales: physical aggression =.85), verbal aggression (= .72), anger (= .83), and hostility (= .77). 3), Interpersonal reactivity index consisting of four 7-item subscales, including Fantasy (FS), Perspective Taking (PT), Empathetic Concern (EC), and Personal Distress (PD), and 4) Self-Construal scale composed of 15 items made up the Independent self-construal subscale, and the remaining 15 items corresponded to the Interdependent self-construal subscale.

2. Database Structure and File Naming

The recordings are arranged in 54 sessions. Each of which consists of one continuous recording of an interaction between two persons.

3. Audio

Three channels of sound were recorded using a MOTU 8pre[1] eight-channel audio interface (see Fig. 1). All audio was recorded at 48kHz sampling and 24 bit per sample. Channel 1 and 2 contain the audio signal from head-worn microphones, type AKG HC 577 L. Channel 1 is for participant 1, and channel 2 for participant 2. Channel 3 contains the audio signal from a AKG C 1000 S MkIII room microphone, hanging from the ceiling, offset from the middle, equidistant to both participants and close to the wall. This channel can be used to obtain a noise estimate for noise reduction in the first two channels. Most of the noise originates from the building ventilation system, which was controlled remotely. Please note that, on some occasions, the ventilation system was shut down or activated during a recording. Therefore, the background noise cannot be assumed constant.

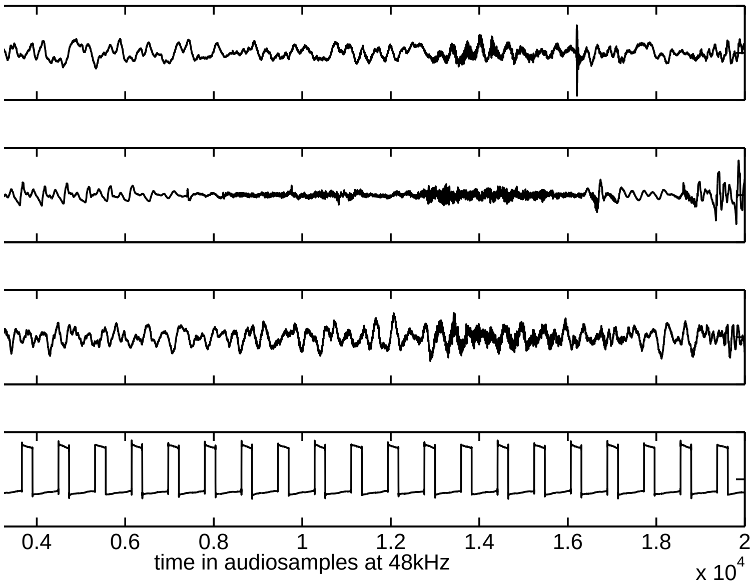

Figure 1. 4 tracks recorded in parallel by MOTU 8pre audio interface. From top to bottom: (1) head microphone participant 1; (2) head microphone participant 2; (3) room microphone; (4) camera trigger with the block pulses indicating the 5ms long frame exposures.

Each session contains two audio files, ending with the post-fix "*_TrimmedAudio_c1_c2.wav" and "*_TrimmedAudio_c3.wav", respectively. The former contains two channels (stereo) with the audio recordings from the head-worn microphone of person 1 (c1) and person 2 (c2) in the left and right channel, respectively. The second audio file contains a single channel (mono) with the audio recording from the room microphone (c3).

The audio files are trimmed such that the start at 1/2 video frame capture period before the centre of the capture time of the first video frame in each of the video

[1] http://www.motu.com/products/motuaudio/8pre

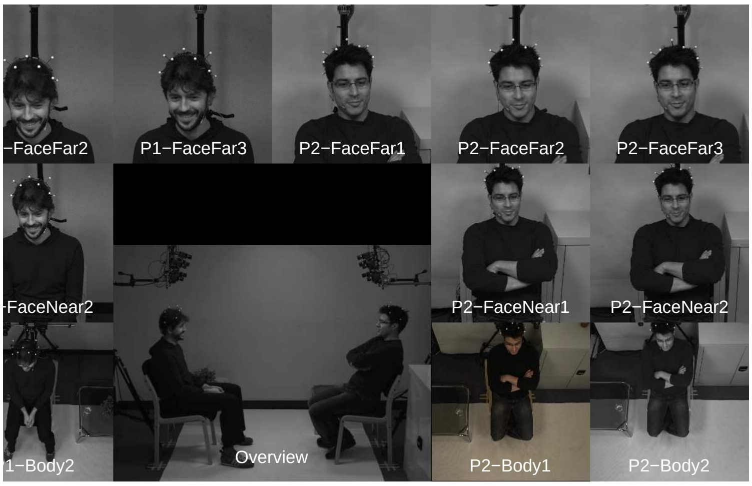

Figure 2. Simultaneous views from all cameras.

sequences. The ending of the audio file is at 1/2 of the frame video frame period after the centre of the capture time of the last video frame.

4. Video

Three types of cameras have been used: An Allied Vision Stingray F046B, monochrome camera, with a spatial resolution of 780x580 pixels; two Prosilica GE1050C colour cameras, with spatial resolutions of 1024x1024 pixels; and 12 Prosilica GE1050 monochrome cameras, with spatial resolutions of 1024x1024 pixels. Different sets of cameras have been set up to record the face regions at two distance ranges: Far corresponds to a distance range for upright poses and Near corresponds to forward- leaning poses. The focal length and focus of the cameras have been optimized for the respective distance range. The best camera view to use for a facial analysis depends on a persons body pose in each moment. The cameras were intrinsically and extrinsically calibrated. See figure 2 for the camera views.

An example of all camera views is shown in figure 2. Table 1 enumerates all camera view names and descriptions. Three types of cameras have been used in the recordings: One Allied Vision Stingray F046B, monochrome camera, with a spatial resolution of 780x580 pixels; Two Prosilica GE1050C colour cameras, with spatial resolutions of 1024x1024 pixels; And 12 Prosilica GE1050 monochrome cameras, with spatial resolutions of 1024x1024 pixels. Different sets of cameras have been set up to record the face regions at two distance ranges: 'Far' corresponds to a distance range for upright poses and 'Near' corresponds to forward-leaning poses. The focal length and focus of the cameras have been optimised for the respective distance range. This means that the best camera view to use for a facial analysis depends on a person's body pose in each moment. The cameras were intrinsically and extrinsically calibrated. The extrinsic calibration is shown in figure 3.

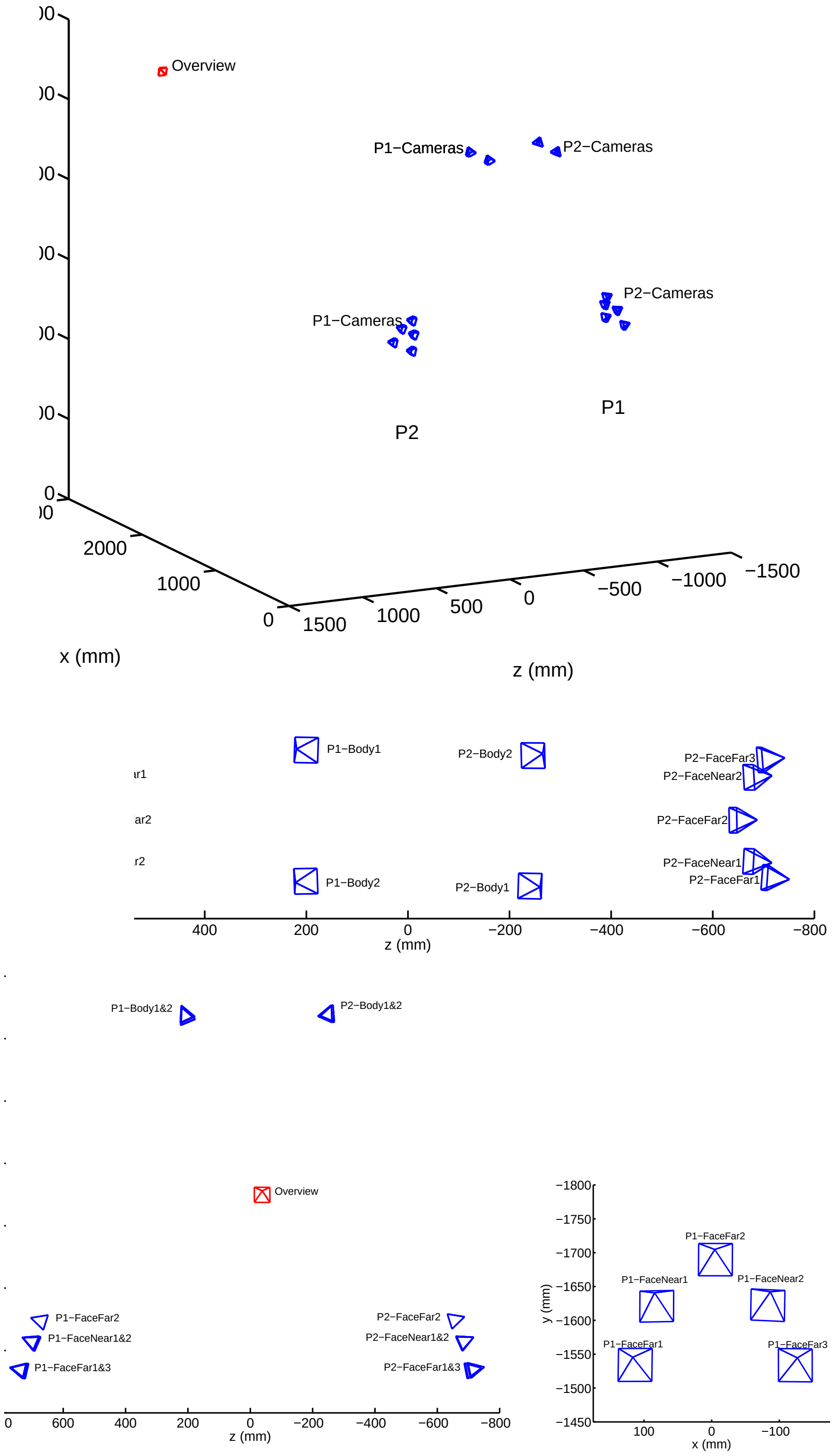

FIGURE 3. Extrinsic camera poses.

Table 1. Intrinsic camera parameters.

cam#	**name**	**description**
1	Overview	Profile view of participant 1 and 2 together. 780x580 pixels monochrome, 5ms exposure.
2	P1-Body1	High view of participant 1, showing the whole body. 1024x1024 pixels colour, 10ms exposure.
3	P1-Body2	High view of participant 1, showing the whole body. 1024x1024 pixels monochrome, 5ms exposure.
4	P1-FaceFar1	Participant 1, close up of face area in upright postures. 1024x1024 pixels monochrome, 5ms exposure.
5	P1-FaceFar2	
6	P1-FaceFar3	
7	P1-FaceNear1	Participant 1, close up of face area in forward-leaning postures. 1024x1024 pixels monochrome, 5ms exposure.
8	P1-FaceNear2	
9	P2-Body1	High view of participant 2, showing the whole body. 1024x1024 pixels colour, 10ms exposure.
10	P2-Body2	High view of participant 2, showing the whole body. 1024x1024 pixels monochrome, 5ms exposure.
11	P2-FaceFar1	Participant 2, close up of face area in upright postures. 1024x1024 pixels monochrome, 5ms exposure.
12	P2-FaceFar2	
13	P2-FaceFar3	
14	P2-FaceNear1	Participant 2, close up of face area in forward-leaning postures. 1024x1024 pixels monochrome, 5ms exposure.
15	P2-FaceNear2	

5. Audio/Video Synchronization

The cameras are synchronised by hardware-triggering, and configured to have exposure intervals around the same center, at 58 frames per second. This is achieved by triggering the monochrome cameras (with 5ms exposure duration) 2.5ms later than the colour cameras (with 10ms exposure duration).

To synchronize between audio and video, we recorded the camera trigger signal as a fourth signal, in parallel with the three audio channels (see Fig. 1). Since the analog inputs of the 8Pre are sampled using the same clock signal, an event in one of the channels can be directly related to a temporal location in all other channels. The camera trigger pulses can be easily detected and matched with all the captured video frames, using their respective frame number and/or time stamp. With the audio sampling rate of 48kHz, the uncertainty of localizing one camera trigger edge is around 20μs. When the 5μs latency and jitter of 10ns of the camera exposure is taken into account, and the timing of multiple trigger pulses is combined, the resulting synchronization error between audio and video can be kept well below 20μs. More details about the synchronization can be found in [1].

6. Annotation

The database has been segmented into speech acts, and annotated for a number of social signalling cues, as well as conscious and nonconscious higher-level behaviours.

6.1. **Segmentation into Episodes of Interest.** In our data, Experiment A includes two parts: presentation and discussion. In the presentation part, it is obvious that interviewees play a role as speakers while the interviewers listen all response from listeners is on the involvement or understanding level. For example, understanding can be expressed by nods. So it is natural that the range of nonverbal behaviour expressed by a listener is small, often limited to cues such as nodding, smiling, and certain mannerisms. On the contrary, in the discussion part, interviewers and interviewees both need to express an actual response, i.e. to give feedback on a communicative level. Even more interesting is that people often only mimic anothers behaviour when they are playing the same role in interactions. In other words: people may not immediately mimic the speakers behaviours while listening, and they may, instead, express a consensus response (since they are functioning on the involvement or understanding level). But when the former listener subsequently takes on the role of speaker, s/he often mimics their counterparts behaviour that was expressed during the previous turn. This complies with one of the most important factors that can affect mimicry - similarity: The similarity of roles played in interactions. In Experiment B, the participants have complete similarity of conversational goal, which is to find a roommate successfully.

In summary, the analysis of relevance among mimicry and social interactions can be extended not only for recognizing human affect, but also for judging relationships (roles) and interaction management (turn-taking).

Annotation Steps:
Segmentation into episodes according to utterance tokens acquired from participants
Annotation of speakers and listeners
Annotation of behavioural cues for both participants separately
Annotation of mimicry

In our annotation tool, options for behavioural cues are predefined. After the annotation of episodes and behavioural cues, the tool can automatically compare whether the selected options are the same for both participants, from which the mimicry label (PRESENT/NOT PRESENT) is derived.

6.2. **Annotation within Segments.** For the episodes of interest, more detailed annotations are included, consisting of behavioural expression labels, mimicry/non mimicry labels, and social signal labels. In the interface of the annotation software, the first item that is provided concerns the behavioural expression labels: smile, head nod, headshake, body leaning away, and body leaning forward. When the video data is played, the annotator has to enter the time when a particular cue was observed, and choose a suitable label from the list. Cases where none of the available labels are appropriate for a certain expression are also taken into other account. Secondly, in order to learn more about the intent behind those behavioural expressions, for each behavioural expression the label of conscious and unconscious is also recorded. For unconscious behaviours, a SOCIAL SIGNAL EXPRESSION has to be chosen. This can be e.g. understanding, agreement, liking, confused, or uncertain. For conscious behaviour, a DESIRED GOAL has to be chosen. For example: to flatter others, to emphasize understanding, to express agreement, to share rapport/empathy, to increase acceptance. Since it is sometimes difficult, or even impossible, to specify a unique reason for mimicry, space is provided to include a comment.

Current annotation considers visual behaviour and participants roles in each conversation. Further annotation will include the participants affect and implied social signals relative to mimicry. It will be mainly based on the questionnaires taken during the experiments.

7. Acknowledgment

This work has been funded in part by the European Communitys 7th Framework Programme [FP7/20072013] under the grant agreement no 231287 (SSPNet). The work of Jeroen Lichtenauer and Maja Pantic is further funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG 203143 (MAHNOB). The work of Michel Valstar is also funded in part by EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching.

References

1. Jeroen Lichtenauer, Jie Shen, Michel Valstar, and Maja Pantic, *Cost-effective solution to synchronised audio-visual data capture using multiple sensors*, Tech. report, Imperial College London, 180 Queen's Gate, London, UK, 2010.
2. X. Sun, J. Lichtenauer, M.F. Valstar, A. Nijholt, and M. Pantic, *A multimodal database for mimicry analysis*, Proceedings of the 4th Bi-Annual International Conference of the HUMAINE Association on Affective Computing and Intelligent Interaction (ACII2011) (Memphis, Tennessee, USA), October 2011.

Imperial College London, Department of Computing
E-mail address: j.lichtenauer@imperial.ac.uk