# A new multichannel multi modal dyadic interaction database

*Viktor Rozgić*[1], *Bo Xiao*[1], *Athanasios Katsamanis*[1]
*Brian Baucom*[2], *Panayiotis G. Georgiou*[1], *Shrikanth Narayanan*[1,2]

[1]Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA
[2]Department of Psychology, University of Southern California, Los Angeles, CA, USA
http://sail.usc.edu[1], baucom@usc.edu[2]

## Abstract

In this work we present a new multi-modal database for analysis of participant behaviors in dyadic interactions. This database contains multiple channels with close- and far-field audio, a high definition camera array and motion capture data. Presence of the motion capture allows precise analysis of the body language low-level descriptors and its comparison with similar descriptors derived from video data. Data is manually labeled by multiple human annotators using psychology-informed guides. This work also presents an initial analysis of approach-avoidance (A-A) behavior. Two sets of annotations are provided, one based on video only and the other obtained by using both the audio and video channels. Additionally, we describe the statistics of interaction descriptors and A-A labels on participants' roles. Finally we provide an analysis of relations between various non-verbal features and approach/avoidance labels.

**Index Terms**: behavioral signal processing, multi-modal database, dyadic interaction, approach and avoidance

## 1. Introduction

Human communication is a dynamic process where communicative goals are achieved through an interactive process employing multi-modal cues: speech and visual cues, including explicit and implicit information such as paralinguistic phenomena and body language. Although complex, the verbal and non-verbal behaviors in dyadic or small group interactions follow patterns that have been the research focus of psychologists for a long time. For example, psychologists have developed many coding schemes, such as the Couple Interaction Rating System (CIRS) [1] and the Rapid Marital Interaction Coding Scheme (RMICS) [2], for annotation of couples interaction and family therapy sessions. These schemes are based on recognition of low-level verbal and non-verbal cues (e.g. gaze, body orientation, turn taking patterns, presence of negative words, tone of voice etc.). Inferences by the experts can be made using these low-level cues towards deriving high-level behavior codes (e.g. acceptance, positivity, blame, negativity, approach-avoidance etc.) with direct influence on evaluation and planning of the therapy process.

Developments in speech (speaker diarization [3]), audio-visual (tracking humans [4], head orientation [5], facial feature extraction [6], hand tracking [7]) and natural language processing have opened avenues of possibilities for automation of the low-level descriptor extraction. An emerging research area, behavior signal processing [8] including social and emotional aspects, focuses on estimation of the high and intermediate level behavior labels from automatically extracted low-level descriptors and design of alternative intermediate level labels that are intuitive and strongly related to the high level labels. For example in emotions research [9], valence and activation are used as an intermediate representation for categorical emotion classification.

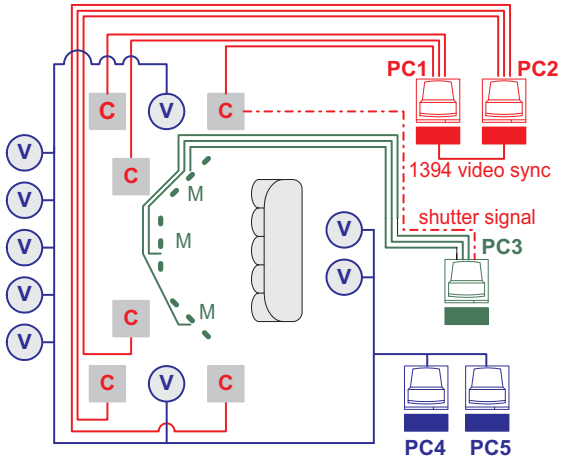


Figure 1: Recording system architecture: blue - video recording system, red - motion capture Vicon system, green - microphone recording system, $PC_1$ and $PC_2$ synchronized via dedicated 1394 bus and synchronized with $PC_3$ using shutter signal

This work provides two contributions. First, we present our multi-modal recording environment aimed at collection and informed analysis of human behaviors in collaboration with psychology experts. We describe the collection and present some initial analysis of the first part of the database of three hours of data consisting of multiple five minute dyadic interactions; a product of the collaboration between the USC Viterbi School of Engineering and the USC Department of Psychology[1]. Each short interaction represents an argument on one of nine suggested topics where each participant is trying to provide evidence that supports her/his point of view. Some of the topics are confrontations about cheating in a relationship, a drinking problem, stealing from a roommate etc. Data is manually transcribed, segmented in speaker turns and annotated by experts with the approach/avoidance labels. The recording environment contains an array of 10-high-definition video cameras, multiple microphone arrays (13 mic total), 2 lapel microphones and a 12-camera motion capture system.

This design allows collection of synchronized high quality signals in a controlled environment and enables investigation of advanced signal processing techniques. For instance,

[1]The described part of the database is collected through role-playing, but we in-parallel analyze real data [10] and intend on collecting real-couple interaction data in this environment in the future.

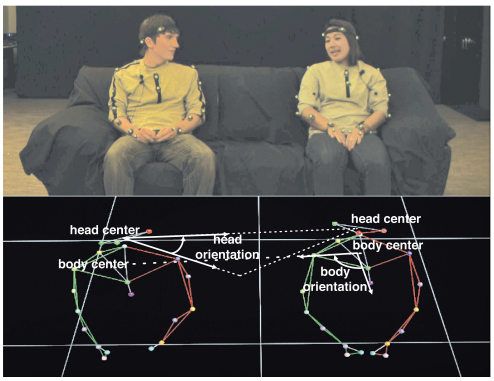


Figure 2: (a) Marker placement: 4 head markers, 3 back markers, 2 chest markers and 7 markers on each arm; (b) Reconstructed marker locations and derived features

the corpus of real married couple interactions used in [10], although at the moment more realistic in terms of the impact to the field of psychology, it restricts the use and development of algorithms. It was not designed and collected to also favor automatic processing and the included audio-video recordings may be of considerably low quality. In addition to these psychology domain data used in [10] our lab has already released an acted multimodal database (http://sail.usc.edu/iemocap) of emotional interactions. We also intend to disseminate to the community this richer in realism and sensing database.

An additional important advantage of the database is availability of both motion capture and video data. This allows us to (a) analyze the relation of the body language features obtained from the motion capture output with domain labels; (b) perform training and testing of algorithms that extract equivalent or similar features from video and (c) analyze information loss through the video processing and refine the algorithms appropriately.

In relation to this, we present the second contribution, statistical dependence of interaction descriptors, e.g., turn durations, number of questions, backchannels, successful and unsuccessful interruptions, on participants' roles and analysis of the relation between various non-verbal features obtained from the audio and the motion capture data and approach/avoidance labels.

Section 3 describes details of the recording environment, the collected database and the annotation process. In Section 4 we present the feature extraction process. Section 4.2 provides statistics of the audio turn taking and motion capture head orientation and hand movement features. In Section 4.3 we present analysis of relation between non-verbal features and the approach/avoidance labels. We conclude in Section 5 with discussion and the future work directions.

## 2. Recording Environment and Hardware

An overview of the recording environment is presented in Fig. 1 and here we provide a description of the hardware setup used for the data collection. Our sensing capabilities include:

- Vicon motion capture system: 12 motion capture cameras that track and record positions of 23 markers on participants' upper bodies (Fig. 2) at 120fps rate. Note that the markers are placed in a way that leaves the skin on the lower arms and face visible allowing algorithm development from the video channel.
- PointGrey camera array: 10 Flea2 PointGrey cameras recording 2 frontal close-up and 8 ceiling far-field views of interaction at 30fps with resolution $1024 * 768$
- Microphones: three 4-microphone T-arrays, a lapel microphone for each participant and a shotgun microphone all recording interaction audio at 48kHz with 24 bit precision

We used 2 PCs with solid state hard drives in RAID 0 configuration to achieve necessary writing speeds. The used configuration supports recording from eight 2Mpix cameras at 30fps in raw format with 8bpp without dropped frames. Due to the huge volume of the recorded data and our processing goals we opted for a resolution of 0.7Mpix per camera-frame. Cameras were synchronized using a dedicated 1394 connection between PCs and PointGrey's Multisync software. We developed the recording $C++$ software using PointGrey's SDK. Audio was recorded using another PC and two daisy-chained 8-channel MOTU-896 devices. The important issue of audio-visual synchronization was addressed by bringing the shutter signal from one of the cameras as an input to the MOTU audio device and recording it synchronously with all audio signals. The synchronization of the motion capture stream with audio-visual components is done using director's clap at the beginning and the end of each recording session. The audio-visual synchronization precision is practically one audio sample, and the synchronization with the motion system is defined by the frame capture rate and is approximately 10ms. A schematic representation of the recording system with connections between different modalities is given in Fig. 1.

## 3. Database

The dyadic multimodal database will include several levels of realism from the human-aspect side. We have initially started our collection with unscripted role-playing scenarios and we intend to continue with more realistic data of couples interacting on conflictual topics of their choosing. We also want to solicit feedback from the broad scientific community and guide the future collection appropriately.

For the first part of the database participants were given time to prepare for arguments on a chosen subset of nine proposed topics and encouraged to be passionate in arguing their position during conversations. Suggested topics were open ended (i.e. couple arguing over fling with a friend or friends arguing over fling with one's boy/girlfriend) and participants were drawing from their own experiences to create a back story that supports their point of view. It was suggested to choose the back stories in a way that makes discussion as natural as possible for the participant. The interactions were segmented, transcribed and labeled with the approach/avoidance labels.

### 3.1. Collection protocol

The data collection protocol contains two main stages. In the first stage, two days before the scheduled collection, participants were given a list of nine scenarios with instructions on how to interact. The second stage happens at the scheduled collection time and represents a sequence of preparation and data collection steps. At the beginning of each session participants were introduced to each other and time was given to them to pick $4-6$ scenarios they would like to discuss during the collection. Participants interact in scenarios that require them to be familiar with each other (couples or friends) so before the recording of each scenario they spend an additional $5-10$min to share information that they considered necessary for the role-play.

After recording all interactions of the same couple we recorded additional reference head orientation data. We also recorded visual and audio information of the environment such as scene and noise backgrounds and data necessary for the joint calibration of all modalities.

### 3.2. Data collection progress

Our data collection is on-going. The first part of the database described here contains approximately 3 hours of data. One third of the interactions contain couples of the opposite sex while the rest involve interlocutors who are of the same sex, mostly female. In these cases participants are acting as friends.

A subset of the data corresponding to eight sessions (45min) is fully annotated and this is the dataset portion we use for the analysis presented in this paper.

### 3.3. Manual post-processing and data annotation

The multi-view motion capture system is designed to track markers on the human body in a 3D coordinate system. However, since the participants were asked to have a natural interaction marker occlusions happen very often. The proprietary Vicon iQ software can not reconstruct reliably (after an occlusion, a wrong label is usually assigned to the occluded marker) full marker trajectories and a manual intervention is necessary. We manually corrected marker labels as needed to enable trajectory reconstruction.

We split the annotation process in two parts. The first part is conducted by labelers who can speak and write English and the second part is conducted by psychology-domain experts trained in coding schemes such as the CIRS. Labels in the first group include all labels derived from audio: speaker segmentation, transcription, dialogue acts on sentence level (question, statement, back-channel) and turn taking labels (successful and unsuccessful interruptions). No labeling of video channel for low-level interaction descriptors was performed since these labels can be extracted directly from the motion capture output. Labels in the second group will include subject-interaction level labels, e.g., presence of blame, attitude (positive vs. negative), acceptance and approach-avoidance, or labels on the sub-interaction level.

### 3.4. Some Illustrative Interaction Statistics

We have extracted a range of features and statistics for all the collected data and description of all is beyond the length constraints of this paper. These include speaker ID based segmentation, speech segmentation using voice activity detection [11], energy, pitch, 13 MFCCs and a microphone array based speaker segmentation [12]. In addition, we calculate a range of functionals, e.g., mean, standard deviation, minimum and maximum. Both for active speaker and inactive participant estimate the number of interruptions and the total interruption duration normalized by the turn length.

Features derived from the motion capture data are chosen according to the approach and avoidance coding manual in a way that intuitively describes relative participant orientation, movement and body posture. For both participants we extract the following functionals in 3sec intervals with 1sec shift: (a) head/body orientation angle relative to the other participant; (b) arm velocity measure representing average velocity of scaled arm markers maximized over left and right hand and (c) measure of how much the body posture is opened/closed in terms of the average distance from left and right forearms from the chest markers.

Based on the best data and annotation we analyzed speaker turn taking, as summarized in Table 1. We can observe that dialogue initiators use longer turns to communicate their messages. As it can be seen from Table 1 initiators use more questions, interrupt the other person more often and according to the number of backchannels they tend to be more active listeners. In addition, our analysis has shown that the turn initiator tends to have significantly more turns of 10 seconds or more while the other person has about 30% more turns of shorter duration.

Table 1: Interaction and dialogue event counts for different interaction roles: Q-question, BC - backchannel, UI/SI - unsuccessful/successful interruption

| session ID | role | Q | BC | UI | SI |
|---|---|---|---|---|---|
| 1,2,3 | initiator | 34 | 11 | 7 | 7 |
| | other | 18 | 1 | 5 | 3 |
| 4,5 | initiator | 11 | 1 | 2 | 4 |
| | other | 7 | 0 | 3 | 0 |
| 6,7,8 | initiator | 32 | 6 | 21 | 27 |
| | other | 4 | 6 | 7 | 18 |
| all | initiator | 77 | 18 | 30 | 38 |
| | other | 29 | 7 | 15 | 21 |

In addition we provide some analysis of the motion capture data as those relate to speaker activity. Fig. 3 represents the velocity of the head and hand motion. As we can see the active speaker demonstrates significantly more movement than the listener. The listener has negligible head and hands movement twice longer than the active speaker, while the active speaker demonstrates a consistently higher velocity of movement.

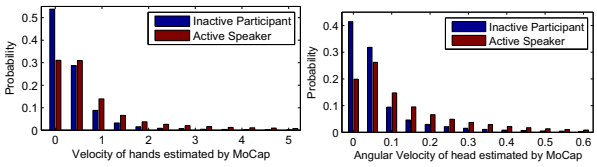


Figure 3: Histogram of probability of the velocity of movement of hands and angular velocity of the head. As we can see the inactive participant is significantly less animated than the active speaker.

## 4. A Case Study: Approach and Avoidance

In this section we present our analysis results of the relation between different low-level descriptors and the approach/avoidance labels on the speaker turn level.

### 4.1. Approach and Avoidance - Expert Annotation

The recorded data represents flow of verbal and non-verbal cues and, in order to avoid apriori interaction segmentation, for the purpose of our analysis experts provided us with the continuous-in-time and discrete-in-value $[-4, -3, \ldots, 4]$ approach-avoidance labels for each participant. Each interaction is annotated by a single expert in two ways: (a) using multi-view video only and (b)using multi-view video with audio. Additionally two test interactions are labeled by three annotators to give an initial insight in annotators agreement. Labels for video only are obtained by following labeling rules related to the gaze, relative inter-participant body and head orientation and qualification of the body pose as opened or closed. Beside visual cues, engagement in the conversation, dialogue management and turn taking behavior were used for the audio-visual labeling.

### 4.2. Interaction descriptors - role dependency

All suggested scenarios have common role profiles, one participant is initiating the discussion with the clear message and goal in mind, e.g., confront friend about her/his drinking problem/cheating or insist on changes in holiday plans, and the other participant is trying to express his point which can end in agreement or participants may stay confronted. We examined dependence of important dialog properties on participants' roles.

Table 2: Approach-avoidance (A-A) label values for different interaction roles

| session ID | role | audio and video | video only |
|---|---|---|---|
| 1,2,3 | initiator | $2.16(0.49)$ | $1.05(0.79)$ |
| | other | $1.88(0.55)$ | $1.39(0.58)$ |
| 4,5 | initiator | $2.66(0.83)$ | $-0.14(0.89)$ |
| | other | $1.39(0.52)$ | $-0.07(0.32)$ |
| 6,7,8 | initiator | $2.25(0.52)$ | $1.73(0.72)$ |
| | other | $1.88(0.48)$ | $1.01(0.58)$ |
| all | initiator | $2.31(0.62)$ | $0.91(0.75)$ |
| | other | $1.78(0.54)$ | $1.14(1.05)$ |

In the Table 2, we present mean and standard deviation for the approach-avoidance labels for each role. From this table it can be seen that different roles were less discriminative in the visual cue domain, while addition of acoustic cues made interaction roles more separable.

### 4.3. Analysis of non-verbal features for A-A estimation

We analyzed the relation of features derived from the motion capture output to the A-A labels derived from video only and combined audio and video. For that purpose we calculated features derived from audio and motion capture output in 3sec intervals with 1sec shift as described in Section 4. In Table 3 we present *mutual information* (MI) values for the chosen set of features. The MI values are estimated by discretizing each feature separately using k-means algorithm with 10 clusters and calculating mutual information between discretized feature variables and discrete A-A labels. The MI is calculated by concatenation of samples (the feature and the label) for all sessions and for all participants.

Table 3: Mutual information between motion capture features and A-A labels

| description | functional | video only | audio and video |
|---|---|---|---|
| body orientation | mean | 0.42 | 0.40 |
| | min | 0.40 | 0.37 |
| | max | 0.47 | 0.37 |
| opened/closed hands *vs* body | mean | 0.45 | 0.27 |
| | min | 0.51 | 0.32 |
| | max | 0.43 | 0.24 |
| hands motion | mean | 0.11 | 0.12 |
| | var | 0.13 | 0.15 |
| pitch | mean | 0.08 | 0.12 |
| | var | 0.07 | 0.12 |
| energy | mean | 0.13 | 0.19 |
| | var | 0.14 | 0.17 |

The measures of how opened/closed is the body posture and of the body orientation angle have the highest relation to the A-A labels. They also show higher MI for A-A labeling from the video only stream (see Table 2), which is expected as this feature is based on motion capture and does not include audio features. Although the acoustic features (pitch, energy) do not exhibit high MI, we can still observe that MI has higher values for the A-A labeling of the audio-visual as opposed to the video only steams. The low MI value for these features implies that alternative set of acoustic functionals should be examined.

## 5. Conclusions and future work

This work describes a novel multi-modal recording environment and a database designed to allow the analysis of various verbal and non-verbal behavioral cues in dyadic interactions. The data offers an opportunity to pursue a range of research questions in human behavior signal processing. We present an illustrative preliminary analysis of dependence between different non-verbal features and approach/avoidance labels. We are currently working on four tasks: (a) annotation of the remaining part of the existing database; (b) estimation of the approach/avoidance labels from audio and motion capture derived features, where this estimation is posed as a classification problem; (c) extraction of head orientation and hand movement features from video using skin detection, background subtraction and multi-view 3D reconstruction algorithms and (d) performance comparison of non-verbal feature sets derived from the motion capture and the video on the approach/avoidance label estimation task.

We hope to receive valuable feedback from the Interspeech community to inform future data collection and analysis.

## 6. Acknowledgments

This research was supported in part by the National Science Foundation and the Viterbi Research Innovation Fund.

## 7. References

[1] C. Heavey, D. Gill, and A. Christensen, *Couples interaction rating system 2 (CIRS2)*, University of California, Los Angeles, 2002. [Online]. Available: http://christensenresearch.psych.ucla.edu/

[2] R. E. Heyman, R. L. Weiss, and J. M. Eddy, "Marital interaction coding system: Revision and empirical evaluation," *Behavioural Research and Therapy*, vol. 33, pp. 737–746., 1995.

[3] G. Friedland, O. Vinyals, Y. Huang, and C. Müller, "Prosodic and other long-term features for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[4] D. Gatica-Perez, G. Lathoud, J. M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 601–616, 2007.

[5] C. Canton-Ferrer, C. Segura, J. R. Casas, M. Pardàs, and J. Hernando, "Audiovisual head orientation estimation with particle filtering in multisensor scenarios," *EURASIP Journal on Advances in Signal Processing*, 2008.

[6] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using gabor feature based boosted classifiers," in *SMC*, 2005.

[7] J. MacCormick and M. Isard, "Partitioned sampling, articulated objects, and interface-quality hand tracking," *Book Series Lecture Notes in Computer Science, Springer*, vol. 1843, pp. 3–19, 2000.

[8] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, pp. 1743–1759, 2009.

[9] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and L. Kessous, "The relevance of feature type for automatic classification of emotional user states: Low level descriptors and functionals," in *Proc. Int'l Conf. on Speech Communication and Technology*, 2007.

[10] M. Black, A. Katsamanis, C.-C. Lee, A. Lammert, B. R. Baucom, A. Christensen, G. G. Georgiou, and S. Narayanan, "Automatic classification of married couples' behavior using audio features," 2010, submitted to Interspeech 2010.

[11] P. K. Ghosh, A. Tsiartas, and S. S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Audio, Speech, and Language Processing*, 2010, accepted.

[12] V. Rozgić, C. Busso, P. G. Georgiou, and S. Narayanan, "Speaker tracking and segmentation with microphone array using mixture particle filter: Improvement of multimodal meeting monitoring system," in *Proc. of Multi Media Signal Processing Conference*, 2007.