The TA2 Database – A Multi-Modal Database From Home Entertainment

Stefan Duffner, Petr Motlicek, and Danil Korchagin

***Abstract*—This paper presents a new database containing high-definition audio and video recordings in a rather unconstrained video-conferencing-like environment. The database consists of recordings of people sitting around a table in two separate rooms communicating and playing online games with each other. Extensive annotation of head positions, voice activity and word transcription has been performed on the dataset, making it especially useful for evaluating automatic speech-recognition, voice activity detection, speaker localisation, multi-face detection and tracking, and other audio-visual analysis algorithms.**

***Index Terms*—High-definition video-conferencing, multi-face tracking, multi-modal database, voice-activity detection**

I. Introduction

The TA2 project (Together Anywhere, Together Anytime) [1] explores how technology can help to nurture family-to-family relationships to break down distance and time barriers. This is something that current technology does not address well: modern media and communications serve individuals best, with phones, computers and electronic devices tending to be user centric and providing individual experiences. Technically, TA2 tries to improve group-to-group communication by making it more natural, improving the image and sound quality, and by giving the users the means to easily participate in a shared activity (such as playing a game) or by sharing pictures or videos. In this context, automatic real-time processing of audio and video (e.g. face tracking and speaker localisation) is required in order to determine how many people are present, who is speaking, and when and where people are speaking.

Several multi-modal databases have been recorded in the past, some of them (e.g. [2], [3]), contain recorded audio and video but only for a single person sitting relatively close to the camera. These databases are mostly used for evaluating person verification algorithms and related topics in the biometrics field. A database that is quite similar to ours is the AMI meeting corpus [4] with over 100 hours of recorded audio from a microphone array and video from several cameras. There is also an annotation provided for part of the data. However, the number and position of the participants in the room are mostly fixed. Also, the recorded scenario represents a formal meeting. Thus, compared to our recordings, this data is much more constrained.

In this paper, we present a database containing 2.6 hours of high-definition audio and video data from two separate rooms, where people communicate via a standard video-conferencing system and play online games with each other. The environment is rather unconstrained and noisy, which makes automatic video and audio analysis challenging, especially when a processing in real-time is required.

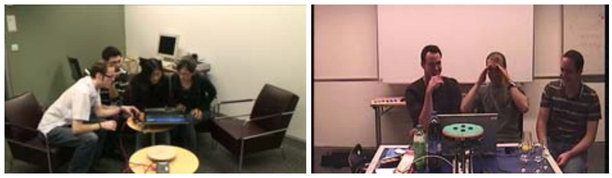


Fig. 1. Example frames of the video recordings. Left: room 1, right: room 2.

Shows two snapshots of the recorded video data from both rooms. Further, manual annotations of head positions and sizes from the video, as well as voice activity and word transcription (with respect to the speaker) from the audio are available in the presented database. All the data can be obtained from the Idiap Research Institute [5].

The paper is organised as follows: in section II, we briefly describe the recording setup. In section III, the recorded data and some statistics are presented. Finally, section IV describes the manually performed annotations.

II. Recording Setup

As mentioned above, the recordings were performed simultaneously in two separate rooms connected via a standard video-conferencing system Fig. 1 illustrates the technical setup and the rough spatial configuration of the devices as seen from the top.

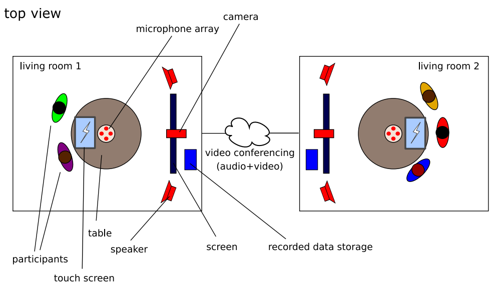


Fig. 2. Layout of the rooms used for the recordings.

The following hardware was used for recordings in room 1:

- Main camera: Sony EV-HD1 (via HD-SDI).
- Capture card: BlackMagic Design DeckLink HD Extreme.
- Video output: HD, 1080i (1920x1080 pixels), 50 fps, converted to 25 fps progressive.
- Microphones: 4x AKG C562CM.

And in room 2, we employed the following setup:

Manuscript received September 20, 2012; revised October 23, 2012.
The authors are with the Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland (e-mail: stefan.duffner@idiap.ch, petr.motlicek@idiap.ch, danil.korchagin@idiap.ch)

- Main camera: Sony SSC-DC58AP.
- Capture card: IVC-4300.
- Video output: 720x576 pixels, 25 fps, progressive.
- Microphones: 8x Sennheiser MKE 2-5-C.

The video-conferencing was performed over an IP network, i.e. separate cameras and microphones from the recording setup were used. The video of the remote party was displayed on the large frontal screen on each side. The electronic (board) game was played on a separate laptop placed on the table between the microphone array and the participants.

Synchronisation of the audio and video was performed manually and offline for each room separately.

III. Data

A. Scenario

The dataset contains one (long) recording session of around 1 hour 20 minutes per each room. The people were free to leave or come in again whenever they wanted. Thus, the number of people changes during the recording, i.e. in the first room there were 3-4 participants and in the second there were 2-3 participants. Two different games were chosen to be played in online mode between both rooms:

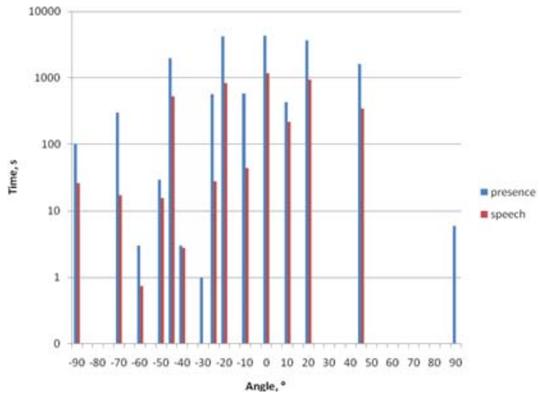
- Battleships (an electronic version implemented in Java).
- Pictionary (using a shared notepad on the screens of the two laptops).

The participants speak in English, but only 1 person can be considered as a native English speaker. There were no constraints on what people should say or do during the recording.

B. Recorded Audio

The audio data was captured by a diamond array with four omni-directional microphones (room 1) and a circular array with eight omni-directional microphones (room 2). It contains an interleaved 4-8 channel Intel PCM audio file (or separate Intel PCM audio files per channel) sampled in 16-bit at 48 kHz. The microphones are numbered counter-clockwise, where the first microphone is pointing to the participants.

In the following, we present some statistics on the data, which illustrate its complexity and the challenge for automatic audio and video processing algorithms. Fig. shows the statistical distribution in logarithmic scale for the overall time of presence at each azimuth in steps of 5° and the overall time of speech coming from the same azimuths, respectively, all extracted from the manual annotation.



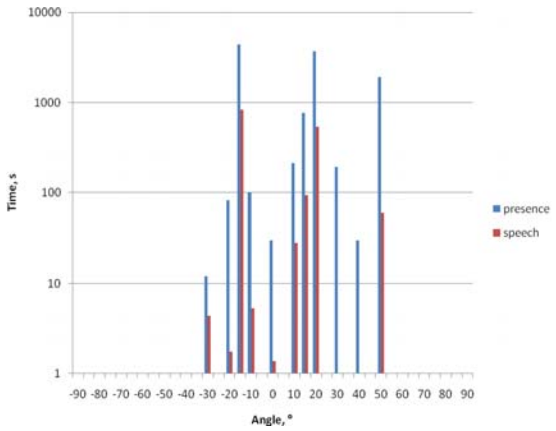


Fig. 3. Time distribution [s] of presence and speech at different angles (Left: for room 1, right: for room 2).

Another statistic is shown in TABLE. It shows the total duration (in seconds) of speech, silence, and cross-talk, for both rooms. One can see that the proportion of cross-talk is considerable, especially in the first room. The Signal-to-Noise Ratio (SNR) is around 17 dB and 26 dB for rooms 1 and 2, respectively, estimated with the method proposed by [6]

Finally, Fig. shows a histogram of durations of individual speech segments. One can see that the majority of utterance durations are below 2 seconds.

TABLE I: Amount of Speech, Silence, and Cross-Talk (in Seconds)

	Room 1	Room 2
Person 1	270.5	691.5
Person 2	440.7	550.4
Person 3	392.3	50.2
Person 4	387.8	190.4
Person 5	0.5	–
Person invisible	33.1	–
Overall speech	1529.9	1482.4
Cross-talk	1001.0	125.4
Silence	2281.2	3132.5
Total length	4806.2	4740.3

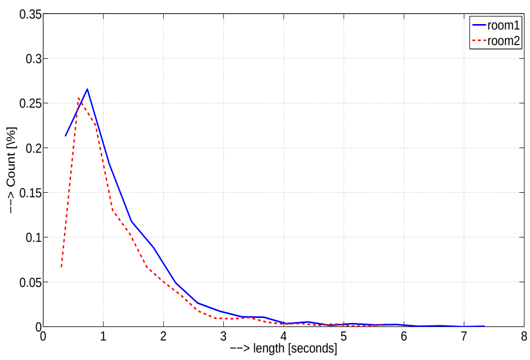


Fig. 4. Histogram of speaking lengths

C. Recorded Video

For room one, the video was recorded and stored at the resolution of 1920x1080 pixels, 25 frames/s (progressive), and encoded in MJPG format.

The video recording of the second room has a resolution of 720x576 pixels at 25 frames/s, and the video format is FMP4 (MPEG4).

IV. Annotation

A. Head Position Annotation

1) Type of annotation

In each (specified) video frame, the locations of all visible heads in the image are annotated. The position and size of each head is described in terms of a bounding box. Moreover, a unique ID (i.e. a number) is assigned to each head/person. That means, if a person leaves the scene and comes back later he/she will have the same ID as before.

To summarise, for a given annotated frame the following information is available:

- number of people present (in front of the camera),
- position and scale of each visible head,
- consistent identities of each visible person,
- the information if a person is occluded or not.

2) Annotated data

Both video files have been annotated. However, not every frame has been annotated. More specifically, the duration between two annotated frames varies from 40 ms (i.e., two consecutive frames) to 10 seconds depending on the dynamics of the video scene (i.e. how much people move). In total, 14550 head positions were annotated.

Head annotation has been performed in the following way:

1) Heads in any pose have been annotated (even if the face is barely visible); exceptions see 4 and 5.
2) The drawn bounding box and the head contour are supposed to coincide.
3) The maximal annotation error is roughly ±15%.
4) Heads that are partially occluded (by other heads, the image border, or other objects) and still visible to at least 50% have been annotated.
5) Heads that are occluded by hands have been annotated.
6) Heads that are fully occluded (i.e. more than 50%) by other heads or objects have been marked as "occluded".

3) Annotation format

The output of the annotation tool is in a specific XML format: The following is the Document Type Definition (DTD) of the output format (file "ta2_annotation_v0_1.dtd"):

```
<!ELEMENT annotation (videoinfo, annotationinfo, personinfo)>

<!ELEMENT videoinfo (name, relpath, length, fps, comments)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT relpath (#PCDATA)>
<!ELEMENT length (#PCDATA)>
<!ELEMENT fps (#PCDATA)>
<!ELEMENT comments (#PCDATA)>
<!ELEMENT annotationinfo (from, to)>

<!ELEMENT personinfo (nbpersons, person*)>
<!ELEMENT nbpersons (#PCDATA)>
<!ELEMENT person (id, sequence*)>
<!ELEMENT id NMTOKEN>
<!ELEMENT sequence (from, to, headposition*)>
<!ATTLIST sequence visibility (visible|occluded|absent) #REQUIRED>
<!ELEMENT from (#PCDATA)>
<!ELEMENT to (#PCDATA)>
<!ELEMENT headposition (time, x, y, w, h)>
```

B. Voice Activity Detection (VAD) annotation

1) Type of annotation

This annotation consists in specifying, for each point in time, if someone is speaking or not, and the identity (i.e. the number) of the respective speaker. In practice, the start and end times for each speech segment as well as the speaker ID has been annotated. Each annotation file contains only the speech annotation of the local speakers (not the ones at the remote location). Remote speech has been marked as "sil" (silence). If a person is speaking somewhere in the room but he/she is not appearing in front of the camera, the speech has been annotated but no ID has been assigned. In this case, the speech segment has been marked with a special flag "speech-inv" (meaning "invisible").

2) Annotated data

All audio files have been annotated. The annotation error of speech/non-speech boundaries is within 50 ms roughly. Very short utterances/sounds (<150ms) as well as very weak sounds have not been annotated.

3) Tools and annotation procedure

For annotating voice activity, the software transcriber has been used [7]. Initial segmentation boundaries have been created automatically and then refined by the annotator, i.e. segment boundaries had to be moved, and segments had to be added or removed. The speaker IDs have also been annotated consistently throughout the respective audio files. These IDs further match the IDs obtained from the video annotation.

4) Annotation format

The voice activity annotation is stored in an XML file. Here is an excerpt of one of the files:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<segments
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    &lt;segment&gt;
        <start>0</start>
        <stop>524</stop>
        <value>speech1</value>
    &lt;/segment&gt;
    &lt;segment&gt;
        <start>524</start>
        <stop>1302</stop>
        <value>speech2 speech3</value>
    &lt;/segment&gt;
</segments>
```

The format is a sequence of <segment> elements with start and stop times specified in milliseconds and the values being "speechN", where "N" is the ID of the speaker, or "speech-inv" if the speaker is invisible (from the camera).

Alternatively, we provide the original annotation file that can be opened and modified with transcriber. This file is also in XML format, but more difficult to process by other software. Here, each segmentation boundary is specified by a "sync" tag

```
<Sync chan="2" time="0.524"/>
S
```

With the time stamp in seconds and the respective channel number followed by the content of the corresponding segment. Only the first audio channel has been used, and the speech segments of different speakers have been marked with "speech1", "speech2", "speech3" etc.

C. Transcription of spoken words and laughter

1) Type of annotation

This annotation consists in specifying which words have been pronounced by the recorded speakers. More precisely, the previously annotated voice activity segments (see Section B) have been merged to the sentence level and then manually transcribed. In addition to words/sentences, segments where a person is laughing have also been annotated and marked as "@laughter".

Noises that are not definable or that are not orally caused by a person are marked as "@noise".

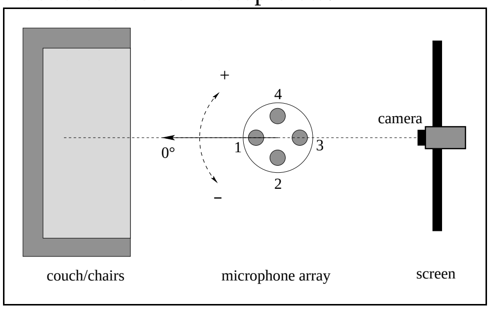
2) Tools and annotation procedure

For obtaining a word transcription, an extension of the software transcriber was employed [8]. The extension handles more than 2 channels (speakers), which is a restriction of the original transcriber. The format is very similar to the original transcriber format (see B.4)).

D. Direction of Arrival (DOA) Annotation

1) Type of annotation

The Direction of Arrival (DOA) of sound (to the microphone array) can be represented as an angle with respect to some reference direction (0°). We define this reference direction as an imaginary arrow intersecting the camera and the centre of the microphone array, facing the participants. DOA angles are then measured clock-wise with respect to the centre of the microphone array. The following diagram illustrates this from a top view:




```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<segments
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    &lt;segment&gt;
        <start>0</start>
        <stop>20000</stop>
        <azimuth id="1">-20</azimuth>
        <azimuth id="2">-0</azimuth>
        <azimuth id="3">+20</azimuth>
    &lt;/segment&gt;
    &lt;segment&gt;
        <start>20000</start>
        <stop>42000</stop>
        <azimuth id="4">-90</azimuth>
        <azimuth id="1">-20</azimuth>
        <azimuth id="2">-0</azimuth>
        <azimuth id="3">+20</azimuth>
    &lt;/segment&gt;
</segments>
```

The annotation of angles is done manually. As no video recording from the top view is available, the annotation represents only a rough estimate. The annotation error is about ±10° and should be considered as an indication of where people are roughly sitting.

2) Annotation format

The DOA annotation is in a specific XML format which can be best explained by an example:

The speakers' locations in terms of angles at one particular point in time are specified inside a <segment> tag. Each speaker's angle is specified with a separate <azimuth> item and an ID, which matches the ID in the corresponding VAD and video annotation files.

V. Conclusion

In this paper, we presented a multi-modal database containing audio-visual recordings of several people located in different (remotely connected) rooms, playing games, and communicating over a video-conferencing system. The selected scenario is rather natural and unconstrained, making these recordings challenging for automatic audio and video processing algorithms. Moreover, extensive manual annotation of audio-visual scene (i.e. head positions, voice activity and pronounced speech provided for each person as well as sound direction of arrival) has been performed. We believe that this work might be very useful for the research community as a reference for evaluating and comparing different automatic audio, visual, or multi-modal analysis algorithms.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme ICT Integrating Project "Together Anywhere, Together Anytime" (TA2, FP7/2007-2013) under grant agreement no. ICT-2007-214793

References

[1] Integrating Project within the European Research Programme 7. Together Anywhere, Together anytime. [Online]. Available: http://www.ta2-project.eu.

[2] E. B. Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J. P. Thiran, "The BANCA database and evaluation protocol," in *Proceedings of the 4th international conference on Audio- and video-based biometric person authentication*, June 09-11, 2003.

[3] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "The Extended M2VTS Database," In *Proceedings of the International Conference on Audio- and Video-Based Person Authentication*, pp. 72-77.

[4] I. M. Cowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus," In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[5] DDP. [Online]. Available: http://www.idiap.ch/dataset/ta2

[6] C. Kim and R. M. Stern, "Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis," *Interspeech*, 2008.

[7] Transcriber. [Online]. Available: http://trans.sourceforge.net

[8] Extensions to Transcriber for Meeting Recorder Transcription. [Online]. Available: http://www.icsi.berkeley.edu/Speech/mr/channeltrans.html