The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements

Élise Arnaud, Heidi Christensen, Yan-Chen Lu, Jon Barker, Vasil Khalidov, Miles Hansard, Bertrand Holveck, Herve Mathieu, Ramya Narasimha, Elise Taillant, et al.

The CAVA Corpus: Synchronised Stereoscopic and Binaural Datasets with Head Movements

E. Arnaud[1,3], H. Christensen[2], Y.C. Lu[2], J. Barker[2], V. Khalidov[3], M. Hansard[3], B. Holveck[3], H. Mathieu[3], R. Narasimha[3], E. Taillant[3], F. Forbes[3], R. Horaud[3]

[1] Université Joseph Fourier, LJK Grenoble, France
[2] Department of Computer Sc Sheffield, United Kingdom
[3] INRIA Rhône-Alpes Montbonnot,France

elise.arnaud@inrialpes.fr, h.christensen@dcs.shef.ac.uk

ABSTRACT

This paper describes the acquisition and content of a new multi-modal database. Some tools for making use of the data streams are also presented. The Computational Audio-Visual Analysis (CAVA) database is a unique collection of three synchronised data streams obtained from a binaural microphone pair, a stereoscopic camera pair and a head tracking device. All recordings are made from the perspective of a person; i.e. what would a human with natural head movements see and hear in a given environment. The database is intended to facilitate research into humans' ability to optimise their multi-modal sensory input and fills a gap by providing data that enables *human centred* audio-visual scene analysis. It also enables 3D localisation using either audio, visual, or audio-visual cues. A total of 50 sessions, with varying degrees of visual and auditory complexity, were recorded. These range from seeing and hearing a single speaker moving in and out of field of view, to moving around a 'cocktail party' style situation, mingling and joining different small groups of people chatting.

Categories and Subject Descriptors

E.m [**Data**]: MISCELLANEOUS; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Sensor Fusion*

General Terms

Design, Experimentation, Human Factors, Verification

Keywords

Binaural Hearing, Stereo Vision, Database

1. INTRODUCTION

Humans' ability to navigate in their environment is extraordinary. We are able to track and recognise moving objects in complex scenes despite attentional distractions such as auditory noise and visual occlusions. These sensorimotor tasks are performed using the limited information supplied by the two eyes and two ears, which is the biological configuration of choice from fish to humans. These sensors, which follow the motion of the observer, sample only part of the scene at any one time.

How this is accomplished is still an open question, and human-centred, audio-visual (AV) scene analysis is a very challenging task. In an attempt to reduce the complexity, supporting research carried out in this area tends to simplify the data and the environment. A lot of work has been done to understand each modality in isolation and binaural hearing and stereoscopic vision research are well-established. However, it is widely understood, that these complementary modalities should be studied jointly. Recent multi-modal research concerned smart room technology and video conferencing systems, which involves multiple, strategically placed cameras and a combination of lapel microphones and microphone arrays. Very few studies limit the sensory input to mimic that of humans both in terms of the number of input channels, and especially in terms of the position and dynamics of the perceiver. This constrats with the growing body of evidence that humans deliberately use movement to improve their sensory input; this might be to move away from distracting noise sources or to reduce occlusion [10]. Certainly evolution has developed sophisticated binaural/binocular fusion strategies that are very poorly understood and barely studied, due to lack of experimental data. An interesting – and under-studied – line of research is to perform an AV analysis of what a person would hear and see while being in a natural environment, and moving the head naturally.

This paper presents a new database, the CAVA corpus, that has been designed to address this central issue and which has recently been made public at [2]. CAVA stands for Computational Audio-Visual Analysis. Such a database is an essential resource to develop tools and test algorithms for an efficient and relevant, human-centred, computational AV analysis of a scene. The database has been established in the framework of the European project POP (Perception on Purpose, FP6-IST-027268). It was recorded in May 2007 by two of the project partners: The University of Sheffield UK and INRIA France. These partners have gathered synchronised auditory and visual data streams recorded using a pair of binaural microphones and a pair of stereoscopic cameras. All data have been recorded from the perspective of an either static or moving/active human or dummy head, and with accompanying head tracking data providing ground-truth information as to the exact head movements associated with the sensory data.

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

The corpus contains recordings from a total of 50 sessions with varying degrees of visual and auditory complexity. These range from seeing and hearing a single speaker moving in and out of field of view, to moving around a 'cocktail party' style situation, mingling and joining different small groups of people chatting. The scenarios were either tailored towards simple, classical AV tasks of tracking the talking head, where either the visual or auditory cues add disambiguating information; or towards more varied environments (e.g. attending a coffee break meeting) with a large amount of rich audio and visual stimuli such as multiple speakers, varying amounts of background noise, occluding objects, faces turned away and becoming obscured and so on. Central to all scenarios is the state of the AV perceiver. As we are particularly interested in exploring the sensory input of an active perceiver, there are recordings where the perceiver is either static, panning or actively moving, mimicking attending to the most 'salient' source at a given moment.

An acquisition device was constructed consisting of a helmet with a stereoscopic camera-pair mounted to the front, and the head tracking target mounted on the top. This helmet was worn either by a person fitted with a pair of in-ear microphones or by a dummy mannequin head equipped with a set of binaural microphones. Further details of the recording equipment are given in Section 3. Section 4 provides details about the post-processing, while Section 5 describes the different AV sequences that are made available to the research community. Section 6 presents some useful tools and illustrates interesting audio / visual properties of the data. Before the technical details, a description of existing AV databases is provided in the following section.

2. EXISTING AV DATABASES

There are several existing databases for the AV research community. In particular, a strong effort has been made to produce a variety of multi-modal databases focusing on faces and speech, like the AV-TIMIT [13], GRID [5], M2VTS [6], XM2VTSDB [16], Banca [9] or CUAVE [17] databases. These databases include individual speakers (AV-TIMIT, GRID, M2VTS, XM2VTSDB, Banca) or both individual speakers and speaker pairs (CUAVE). All have been acquired with one fixed camera and one fixed microphone. A corpus more closely related to ours is the AV16.3 database [15]. It includes a range of situations, from *meeting situations* where speakers are seated most of the time, to *motion situations*, where speakers are moving most of the time. The number of speakers may vary over time. Three fixed cameras were used such that their combined fields of views cover the scene. Two fixed 8-microphone circular arrays were used. The AMI and AMIDA EU projects have produced large collections of fully annotated meeting room based data. In addition to using circular microphone array and multiple cameras, there are also some sessions where binaural data have been recorded using a dummy head/torso. However, because of the physical recording setup, the acoustic quality of these channels is limited, the dummy is static and no corresponding video exists [1]. In terms of evaluation efforts, the CLEAR evaluation and workshop series is closely related [3]. That work is concerned with the evaluation of systems for tracking in 2D and 3D and aims to bring together researchers with similar interests in order to establish a common, international effort.

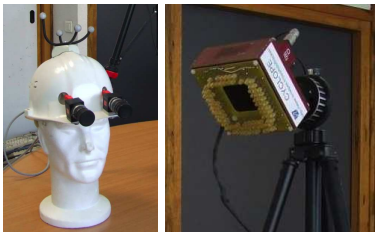


Figure 1: The acquisition device.

The main distinction of our corpus with respect to existing ones, is that it has been designed to model what the AV perception of the scene would be from a human point of view while limiting the recording equipment to a pair of binaural microphones and a pair of stereoscopic cameras.

3. THE ACQUISITION SYSTEM

The entire CAVA database was recorded using the acquisition device depicted in Figure 1. It is comprised of a helmet equipped with a pair of stereoscopic cameras and the CYCLOPE head tracker [4]. The binaural microphones are fitted in the ears of the human or dummy head wearing the helmet. The head tracking device consists of a target placed on top of the helmet to ensure maximum visibility from the head tracking camera, which was mounted in the ceiling during recordings. The helmet was constructed to give a human-like geometric correspondence between "eyes" and "ears" data while allowing the person wearing the helmet to be able to see naturally. The full setup of the recording system is illustrated in Figure 2. The perceiver wore the acquisition helmet and for each scenario, the following five streams were recorded:

- two auditory streams each recorded at 44.1 kHz,
- two visual streams, giving stereo image pairs; 1024x768 at 25 frames per second,
- one tracking stream, providing head position and orientation with 6 degrees of freedom.

The CAVA corpus was recorded in a 7m×5m office-like room with carpets, painted walls and board ceilings. In addition to the fluorescent lamps in the room, two 500 watt studio lamps with light reflectors were used. To minimise unwanted acoustic noise, all computers were positioned outside the room, and all wires run under a door, which was closed at recording times. Figure 3 shows four photographs from the room depicting parts of the setup and scenario sessions. Each element of this setup is described below.

The audio system Different types of microphones were used depending on whether the helmet was being worn by the mannequin or by the human participant. In the case of using the mannequin the Brüel & Kjær (B&K) type 4128C head and torso simulator were fitted with high-specification microphones: two B&K type 4190 half-inch microphones, each connected to a B&K type 2669 preamplifier and in turn attached to a B&K type 2690-0S2 Nexus conditioning amplifier. In the case of a person, we used a microphone earbud head-set (Soundman OKM in-ear binaural microphones connected to a Soundman amplifier). In both cases the signals were captured by a laptop computer using an M-Audio Pro USB A-to-D converter. The audio acquisition is done

Figure 3: Photographs of the CAVA acquisition room and devices.

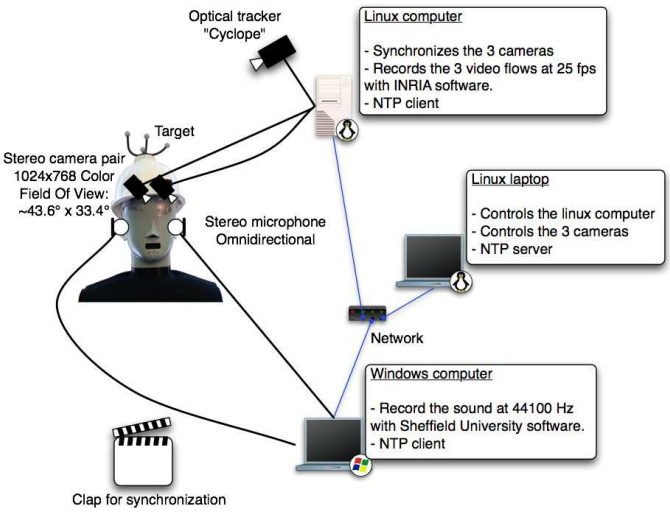


Figure 2: CAVA database acquisition setup.

with in-house software from Sheffield University. The audio stream was sampled at 44.1kHz and a time stamp was saved in the header of the recorded data.

The video system The stereoscopic system is composed of two Point Grey Research Flea cameras fixed on the helmet visor. The two cameras are firewire IIDC compliant, 1024×768 colour cameras with 6mm Fujinon lenses. The recorded data is in raw format, which means 8 bits per pixel with Bayer pattern encoding. This representation is decoded in the post-processing. The video acquisition is done with an INRIA software bundle, that is used to control and setup the cameras, to record the video stream, and to control the recorded data. One file per camera per sequence is recorded and a timestamp is assigned to every frame.

The head tracking system We used the CYCLOPE head tracking system to provide the ground truth for the head movements [4]. The CYCLOPE tracking system is a 6 degrees of freedom (6DOF) input device for virtual reality experiments developed at INRIA. It is composed of two parts, a 640x480 monochrome camera combined with an infra-red flash and a target using spherical retro-reflecting markers. The target used for the CAVA recordings consists of four non-coplanar markers and it was attached to the top of the acquisition helmet enabling the capturing of movement of the perceiver's head. The head tracking camera is fixed to the ceiling of the room positioning it above the scene and therefore always able to see the target. The position of the markers is processed to estimate the coordinates of the target. The resulting coordinates are in the frame of the tracker's camera, which means that the data are interpreted relatively to its initial position.

4. POST-PROCESSING OF DATA

Synchronisation The three video streams (two cameras plus the tracker camera) are timestamped on one single computer, which facilitates the time synchronisation. Furthermore, we used a trigger device so that the images were taken simultaneously on all the cameras. The framerate was set to 25 fps. To synchronise the audio and video streams, we used a simple "clapboard" system, that is easily detected in the audio recordings, and in the video frames.

Calibration Different calibrations are needed in order to use the recorded data. For the video stream, the stereo camera system is calibrated in order to calculate the relative position and orientation of the cameras and process the stereoscopic data. This gives us the intrinsic parameters of the stereo camera pair, as well as external parameters of the cameras. The intrinsic parameters of the cameras may be used to undistort and rectify the images. The external parameters of the cameras are needed to interpret one camera's data from the other's point of view. The calibration process used is the one provided by the image processing library OpenCV [7]. It uses an image sequence of a 'chessboard' at different positions and angles, recorded by the two cameras simultaneously.

The audio files are calibrated through the use of calibration files to ascertain the exact amplification in the left and right ear channel respectively. At certain intervals throughout the recordings, audio calibration files were recorded by attaching a B&K pure tone generator to each of the dummy ears in turn. All following sessions are post-processed with normalisation factors according to the RMS value observed in the calibration files immediately preceding it.

For the integration of AV observations, the microphone positions in the visual frame are needed. An optimization procedure estimates this space mapping from matched observation clusters in the two feature spaces. The CYCLOPE tracker has been calibrated to provide the internal parameters of the couple camera/lens, which are subsequently used for processing.

5. THE CAVA DATASET

A total of 50 sessions were recorded, each one lasting between 20 seconds and 3 minutes. The system setup allowed for the acquisition helmet to be worn either by the dummy head/torso or by a person. The dummy head configura-

tion was used for scenarios requiring little or limited movement. For scenarios requiring more active and human-like behaviour we used a human subject; in practice, the subject movement was limited to the range of the tracking camera mounted in the ceiling, but by ensuring that all the scenario "actors" were themselves very mobile, a high level of activity in the environment was possible.

Table 1 gives an overview of the recordings; the sessions are logically divided into (i) fixed perceiver sessions, (ii) panning perceiver sessions and (iii) moving perceiver sessions. The name of each sequence is unique, and is composed of a scenario name and a number e.g. tracking test one speaker, sequence 1 (TTOS 1). Each scenario has been recorded several times. One representative sequence per scenario is currently available. The names used in the table correspond to the names of the sequence on the web site. Stereo pair images from four sequences are shown in Figure 4.

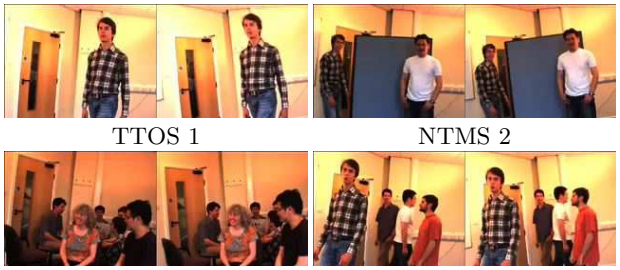


Figure 4: Stereo pairs acquired from four sequences.

5.1 Scenarios with a fixed perceiver

The aim for the fixed perceiver scenarios is to enable evaluation of audio, video and AV tracking and clustering in scenarios with various challenges such as speakers walking in and out of field of view, walking behind a wall, speakers changing appearance and multiple, simultaneous sound sources. These are covered by the following scenarios. Accompanying "storyboard schematics" are given in Figure 5.

TTOS: Tracking test; one speaker - Figure 5(a). One speaker, walking while speaking continuously though the whole scene. The speaker moves in front of the camera and passes behind. He reappears from the right, and turns to the cameras.

CT1OS: Clustering test 1; one speaker - Figure 5(b). One speaker, walking. The speaker moves while speaking in front of the camera and passes behind it from the left. As soon as he gets out of the field of view, the actor becomes silent. Only on reappearing from the right, does he start speaking again and turns to the cameras.

CT2OS: Clustering test 2; one speaker. Same scenario as CT1OS again with one walking speaker. The main distinction is that, when reappearing, the actor has changed appearance (taken off jacket, put on glasses).

CT3OS: Clustering test 3; one speaker - Figure 5(c). Two actors, only one seen and heard at a time. The first speaker moves towards the camera then disappears from the field of view and stops talking. The second speaker enters the field of view while speaking and faces the cameras.

NTOS: Noise test; one speaker - Figure 5(d). One speaker, walking. The actor walks behind a wall and returns to his initial position, always speaking. Various audio noises like clicks and music are present. The lighting condition is intentionally modified.

DCMS: Dynamic changes; multiple speakers - Figure 5(e). Five actors in total. Initially there are two speakers, then a third joins, one leaves, and later on a fifth joins. Then another two leaves. All actors speak and move around.

TTMS: Tracking test; multiple speakers - Figure 5(f). A more complex tracking scenario than the single speaker TTOS. Four actors are initially in the scene. As they start speaking (and go on speaking throughout the test), they move around; one person exits the scene, walks behind the camera while talking, and reappears.

CTMS: Clustering test; multiple speaker - Figure 5(g). A more complex clustering test scenario than the single speaker CTOS. Here four actors are initially in the scene. As they start speaking and moving around, two people exit the scene, stop talking, reappear and start talking again.

NTMS: Noise test; multiple speakers - Figure 5(h). Similar to the one speaker noise test, NTOS. Two speakers are talking, occasionally walking behind a screen.Meanwhile music and clicks are heard in the background.

M1: Meeting - Figure 5(i). Five actors are seated around a table, three are visible to the fixed perceiver (dummy head); one is to the left and one is to the right of the dummy. Initially all join into the same conversation and later on two sub-groups of conversations are formed.

CPP: Cocktail party problem - Figure 5(j). 7 actors in total, 6 in scene and one to the left of the fixed perceiver. Two groups of conversation (one immediately in front of and one further away from the dummy head) are formed. People are seated. At some point one speaker from the most distant group gets up and joins the conversation of the front group. This setup makes for a very challenging auditory and visual scene.

5.2 Scenarios with a panning perceiver

The panning perceiver scenarios were constructed to obtain recordings of controlled cues from an actively moving head. They are all recorded using the dummy head and torso strapped onto a swivel chair. During recordings, the chair is panned from side to side at the same time as the scenario is "acted" out. The following sessions are public:

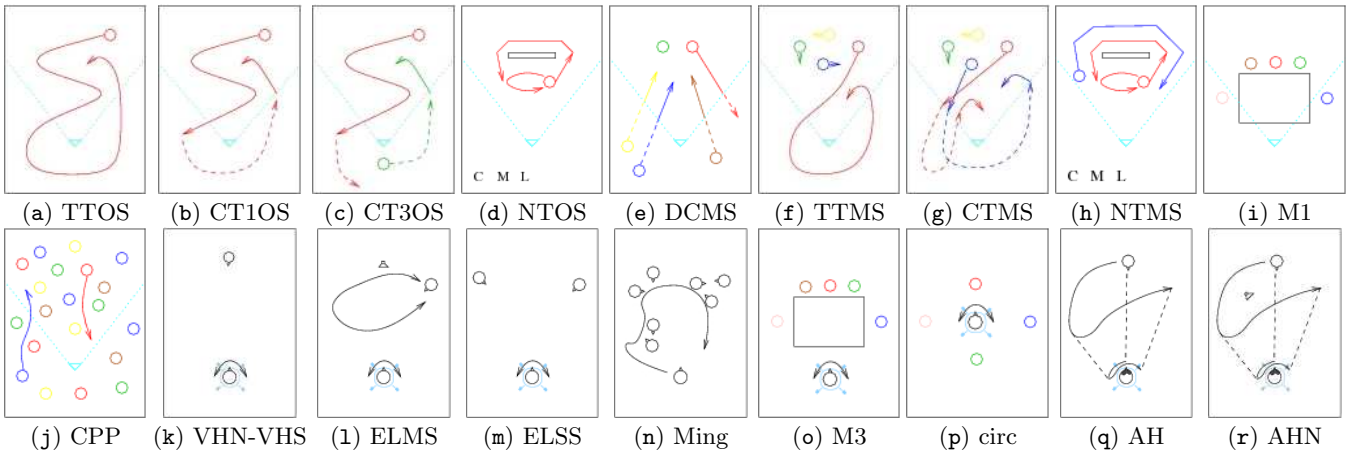
VHS: Varying head speech - Figure 5(k). A single speaker, static while speaking, is standing at 0° azimuth relative to the perceivers' start position. The perceiver starts facing the sound source, and then is moving with periodic left-right movements. The purpose of this sequence is to measure the effects of the perceiver's head movements on binaural cues for localisation

VHN: Varying head noise - Figure 5(k). Similar to the VHS, but with a speaker playing white noise at 0° azimuth relative to the perceivers' start position. The perceiver begins facing the sound source, and then moves from side to side. This will act as a comparison to the cues obtained from speech from the VHS scenario.

ELMS: Ego location moving speech - Figure 5(l). One moving speaker is walking in the scene and may disappear from the field of view. Two additional static sound sources

	sequence name	duration min:sec	type of head	number of speaker(s)	speaker(s)/noise behaviour	visual occlusion	auditory overlap
fixed perceiver	TTOS 1	00:20	dummy	1	moving	yes	no
	CT1OS 1	00:18	dummy	1	moving	no	no
	CT2OS 3	00:21	dummy	1 (changing appearance)	moving	no	no
	CT3OS 1	00:19	dummy	2 (one at a time)	moving	no	no
	NTOS 2	00:33	dummy	1	moving	yes - L	no - M/N/C
	TTMS 3	00:23	dummy	3 to 4	moving	yes	yes
	CTMS 3	00:25	dummy	1 to 3	moving	yes	yes
	DCMS 3	00:48	dummy	2 to 4	moving	yes	yes
	NTMS 2	00:26	dummy	2	moving	yes - L	no - M/N/C
	CPP 1	02:40	dummy	several	seated	yes	yes
	M 1	03:47	dummy	5	seated	yes (2 not seen)	yes
panning perceiver	VHS 1	00:34	dummy	1	fixed	yes	no
	VHN 1	00:32	dummy	1	fixed	no	no
	ELMS 3	00:43	dummy	1	moving	yes	no - N
	ELSS 1	00:41	dummy	2	fixed	yes	yes
	ELSN 1	00:53	dummy	2	fixed	no	yes
active perceiver	AH 2	00:38	human	1	moving	yes	no
	Ming 2	01:44	human	several	moving	yes	yes
	M 3	03:54	human	5	seated	yes	yes
	Circ 1	01:00	human	5	seated	yes	no
	AHN 1	00:44	human	1	moving	yes	no - N
	AHN 4	00:56	human	1	moving	no	no
	P 1	01:17	human	5	seated	yes	no

Table 1: List of recorded sequences. Visual occlusion means either (i) an occlusion of a speaker by another speaker or by a wall, or (ii) a speaker outside the field of view while speaking. In the column "auditory overlap" and "visual occlusion", the tags mean [M]usic, [C]licks, white [N]oise and [L]ight changes.



(a) TTOS (b) CT1OS (c) CT3OS (d) NTOS (e) DCMS (f) TTMS (g) CTMS (h) NTMS (i) M1

(j) CPP (k) VHN-VHS (l) ELMS (m) ELSS (n) Ming (o) M3 (p) circ (q) AH (r) AHN

Figure 5: Scenario schematics. Scenarios (a)-(j) are recorded using a fixed perceiver (dummy head), scenarios (k)-(m) uses a panning perceiver (dummy head on swivel chair) and scenarios (n)-(r) are recorded with a human wearing the helmet and in-ear microphones. Lines indicate actor and perceiver 2D trajectories in the room. A full line indicates "speaking while walking", and a dashed line mean "quiet while walking". When fixed, the field of view is drawn in blue. Actors are shown as circles, and stationary sound sources as triangles. The elongated rectangle represents an occluding wall. The tags mean [M]usic, [C]licks, [L]ight changes.

playing white noise are placed at $-15°$ and $45°$ azimuth relative to the perceivers' start position. The perceiver is static during the first part of the sequence and then is panning randomly. The purpose of this sequence is to study whether the head movement can be inferred from binaural dynamics, and to highlight complementarity between audio and video modalities.

ELSN: Ego location; stationary noises. Two loud speakers playing white noise are positioned at $-15°$ and $45°$ azimuth respectively relative to the perceivers' start position. The dummy is moved in a "random" fashion from side to side. This scenario provides interesting data for investigating to what degree it is possible to do ego-orientation using audio, video and AV cures.

ELSS: Ego location; stationary speech - Figure 5(m). Like ELSN but with speaking people positioned at -15 and 45 azimuth and with a pseudo-random movement of the perceiver.

5.3 Scenarios with a moving perceiver

The aim of the scenarios with a moving perceiver is to provide very challenging AV situations that mimic sitations that can appear in a real-life environment.

Ming: Mingling - Figure 5(n). Small groups of people chatting are placed around the person wearing the acquisition helmet. He turns around to join in different conversations through the scenario. This is a very challenging scene with lots of people talking from all directions and with people moving in and out of the field of view.

M3: Meeting - Figure 5(o). Five speakers are sitting around a table having a conversation with normal turn taking. The sequence starts out with each speaker in turn saying their name and affiliation. The perceiver person is instructed to move naturally in terms of who to face during the conversation.

Circ: Circle - Figure 5(p). Several speakers positioned in a circle around the perceiver. Each person takes it in turn to speak and the person wearing the acquisition helmet attempts to look at them as soon as possible after they've started speaking.

AH: Active hearing - Figure 5(q). One speaker talking while slowly moving around the room following an unpredictable pattern. Blindfolded, seated on a swivel chair, the listener is asked to stay facing towards the moving speaker. The research question that may be addressed with this sequence concerns a behavioural study, i.e. how does the perceiver move in order to track the sound source?

AHN: Active hearing noise - Figure 5(r). The same scenario as AH, with one moving speaker and an additional static, stationary white sound source. The research purpose is similar to the previous scenario, here in a challenging noisy environment.

P: Panel. This scenario is to mimic a person listening into a conversation with non-overlapping speech and looking at the person speaking. Five people are positioned around a table and taking it in turn to count out loud.

6. TOOLS FOR DATA EXPLOITATION

In this section, we present some useful tools, and illustrate some low level A/V cues that may be used to exploit the data. For an example of an AV localisation algorithm tested on the CAVA database, we refer to [14].

Figure 6: Visual cues: (a) original image pair with superimposed epipolar lines and (b) rectified stereo pair

6.1 Visual tools and cues

Rectification of stereo pairs The stereo cameras are fixed on the helmet, approximately parallel to each other, at a separation of 107mm. As can be seen from Figure 6(a), the resulting images are not in epipolar alignment. In particular, there is a large cyclo-rotation. To calculate the relative position and orientation of the cameras, a full metric calibration of the stereo rig was performed. For each stereo-pair, the epipolar geometry may be obtained from the calibration parameters [12]. As an illustration, epipolar lines are depicted in figure 6(a). Finally, the OpenCV library is able to provide the matrices to rectify the image pairs (see figure 6(b)). Such a rectification may be useful for some visual application such a dense stereo estimation. The list of parameters and matrices that may be useful for a CAVA user, i.e. calibration and rectification matrices, are provided along with the CAVA sequences.

Extraction and matching of interest points To provide a reference model of the scene geometry, it is useful, given a binocular pair, to compute a sparse set of features in each images and to estimate the correspondence between them. These low levels cues, also called *interest points*, can be computed using a standard operator [11]. Those points are located at positions of the image where the luminance signal is distinctive such as at corner locations. An example set of interest points in left and right views are shown in Figure 7. Knowing the calibration parameters, and having matched corresponding interest-points in the left and right views [8], their 3D position can be recovered. As shown in Figure 7(c), this gives us some relevant information on the scene geometry. The two connected spheres at the bottom of the picture correspond to the camera pair. The various ellipsoids are obtained with a simple EM clustering algorithm. The three central ellipsoids correspond to the speakers. The top and bottom ellipsoids correspond to the background and to mis-matched images features, respectively.

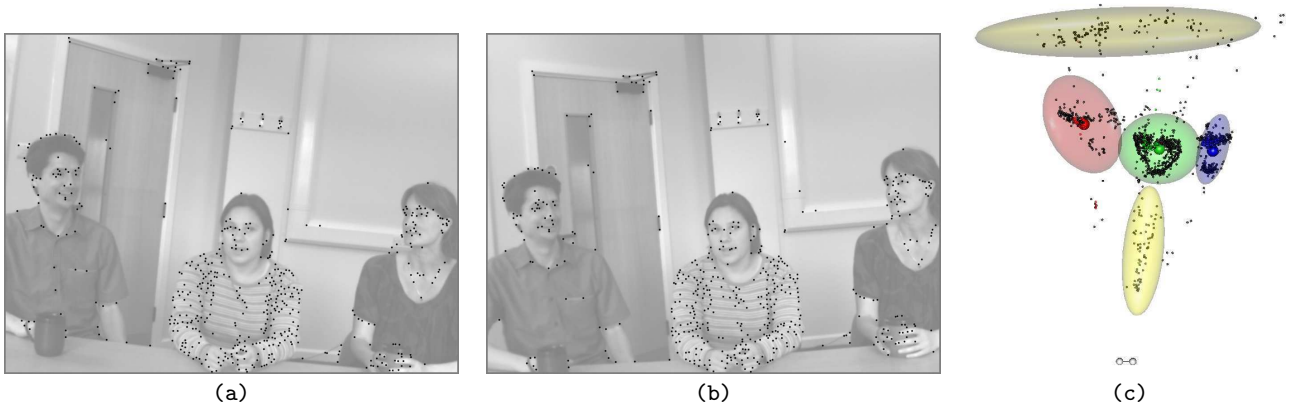


Figure 7: Visual cues: detected interest points in the (a) left and (b) right images; (c) top view of the 3D reconstruction. The two connected spheres at the bottom of the picture correspond to the camera pair. The various ellipsoids are obtained with a simple EM clustering algorithm. The three central ellipsoids correspond to the speakers. The top and bottom ellipsoids correspond to the background and to mis-matched images features, respectively.

6.2 Audio cues

Extraction of binaural localisation cues The two main cues for doing binaural, angular localisation are interaural time diffences (ITDs) and interaural level differences (ILDs) caused by the signal arriving at the ear the closest to the sound source slightly sooner and louder than at the opposite ear.

Figure 8 shows ITDs and ILDs for two varying head sequences: VHS and VHN. The ITDs are computed by the standard procedure of identifying peaks in the cross-correlogram of the left and right ear signal and the ILDs are level differences. The three panels in part (a) show the ITDs, ILDs and the envelopes of the left and right ear signal plotted over time for a about 15 sec. of the VHS sequence. A similar sized segment of the VHN has been used to generate the plots in part (b). Both types of cues estimations are far more accurate for the white noise as opposed to the speech. ITDs are a more reliable cue for the low-frequency bands and for the speech case. The ILDs only provide relatively poor azimuth evidence. This is different for the white noise case, which is more broadband and both ILDs and ITDs are reliable cues. Figure 9 shows a histogram of ITDs for the M1 sequence. Six peaks are clearly visible in the histogram, corresponding to the five speakers in the session as well as to the person operating the clap board and then walking out towards the right of the perceiver. Figure 10 shows ITDs plotted over time for four different sequences: M1, TTOS, ELSS and ELSN. The ITD estimates for the M1 sequence are very noisy with many outliers indicating the complexity of the task. However, all speakers are seated and hence relatively still, so it is possible to detect some short single-speaker segments in the plots. The ITD estimates for the TTOS, ELSS and ELSN sequences are far more sharp. In TTOS the speaker is moving and the trajectory is evident[1]. The ELSN and ELSS sequences were recorded with a panning receiver and involve a moving speaker against a static background of either a white noise source (ELSN) or another speaker (ELSS). Because the standard ITD estima-

[1]Note that from between approximately 18 sec. and 22 sec. the speaker is behind the perceiver, but ITDs are mapped to the front hemisphere on the plot

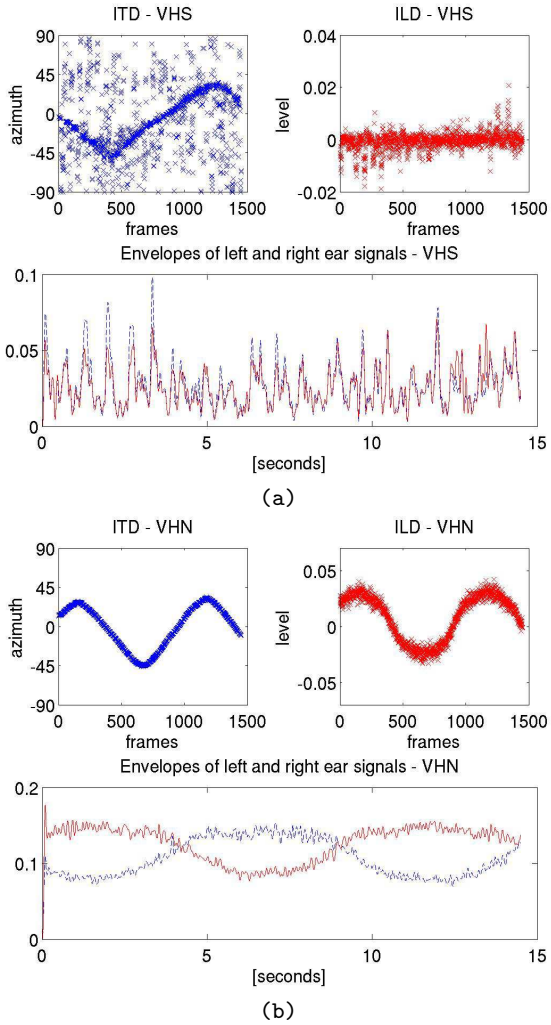


Figure 8: ITDs, ILDs and left and right ear envelope signal for the VHS (a) and the VHN (b) sequence.

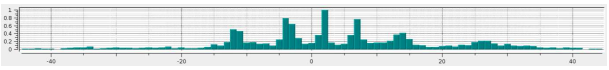


Figure 9: Audio cues: Histogram of ITDs gathered on the M1 sequence. There are five peaks that correspond to speakers, and one (at the right) that corresponds to the clap-board aperator.

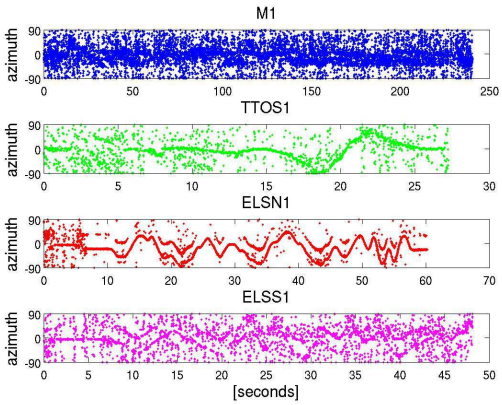


Figure 10: Auditory cues: plots of ITDs over time for four different sequences.

tion, as used here, is restricted to only finding the ITD of the dominant (loudest) source, the trajectories of the two sound sources are not perfect. However, especially for the ELSN sequence, there are clearly two trajectories: one for the "background" which is modulated with the head movement and a second trajectory, which is affected by both the perceiver's panning and the speakers' movement.

6.3 Head tracking data

The head tracking system provides valuable ground-truth data about the relative location of the acquisition device and allows for statistical analysis of head movements and sensory cues. Figure 11 shows a plot of the 3D trajectories as obtained from the mingling session.

7. CONCLUSION

We have presented a new database, captured by an acquisition system that enables the synchronised recording of three streams: one from a pair of binaural microphones, one from a pair of stereoscopic cameras and one from a head tracking system. The data has been made public and it fills a gap in the study of human-centred, audio-visual scene analysis. The CAVA database offers the research community the opportunity to test their multi-modal fusion algorithms on a large, realistic and challenging database. It is hoped that this database will become a standard resource, which will allow institutions to easily evaluate the performance of their algorithms.

8. REFERENCES

[1] Ami project. http://www.amiproject.org/.

[2] CAVA database. http://perception.inrialpes.fr/CAVA_Dataset/.

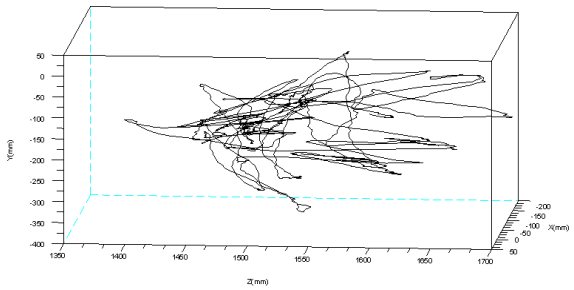


Figure 11: Head tracking trajectory in 3D space, for the mingling session (Ming2).

[3] CLEAR evaluation and workshop. http://www.clear-evaluation.org/.

[4] CYCLOPE tracker. http://www.inrialpes.fr/sed/6doftracker/.

[5] GRID audiovisual sentence corpus. http://www.dcs.shef.ac.uk/spandh/gridcorpus/.

[6] M2vts database. http://www.tele.ucl.ac.be/PROJECTS/M2VTS/.

[7] OpenCV library. http://opencvlibrary.sourceforge.net.

[8] P. Aschwanden and W. Guggenbül. Experimental results from a comparative study on correlation type registration algorithms. In Förstner and Ruwiedel, editors, *Robust computer vision: Quality of Vision Algorithms*, pages 268–282. Wichmann, 1992.

[9] E. Bailly-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 *AVBPA*, 2003.

[10] M. P. Cooke, Y.-C. Lu, Y. Lu, , and R. Horaud. Active hearing, active speaking. In *Int. Symp. Audiological and Auditory Research*, Helsingor, Denmark, 2007.

[11] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Conference*, pages 147–152, August 1988.

[12] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[13] T. Hazen, E. Saenko, C. La, and J. Glass. A segment-based audio-visual speech recognizer: Data collection, development and initial experiments. In *ICMI*, 2004.

[14] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. Horaud. Detection ad localization of 3d audio-visual objects using unsupervised clustering. In *ICMI*, 2008.

[15] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez. Av16.3: an audio-visual corpus for speaker localization and tracking. In *MLMI*, 2004.

[16] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *AVBPA*, 1999.

[17] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *ICASSP*, 2002.