



Inesperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.))

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D2.1 Data Recorded from Different Acoustic Environments

Date of deliverable: 30.09.2006
Actual submission date: 25.01.2007

Start date of project: 01.01.2006

Duration: 60 months

Organization name of lead contractor for this deliverable: **Carl von Ossietzky
Universität Oldenburg**

Revision [2]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other program participants (including the Commission Services)	X
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	



Inesperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.))

D2.1 DATA RECORDED FROM DIFFERENT ACOUSTIC ENVIRONMENTS

Carl von Ossietzky Universität Oldenburg

Abstract:

DIRAC's audio sensors, body-mounted microphone arrays, have been used to capture a diverse body of sound data from different acoustic environments. By recording room impulse responses of the microphone array in addition to actual sound signals, we are able to compose carefully controlled benchmark data for development and evaluation of algorithms in WP1 and WP2. In total, audio recordings of 262 min. length and 308 impulse response sets have been recorded. The fully portable audio recording setup developed for this task has also been combined with the panoramic camera setup of Tomas Pajdla's group (DIRAC partner CZ-CTU) to capture 7 min. of pilot audio-visual data.

Table of Content

1	Introduction	4
2	Recording Setups and Measurement Methods.....	4
2.1	Artificial Receiver.....	4
2.2	Human Receiver.....	5
2.3	Impulse Response Measurement Methods.....	6
3	Description of Recording Situations and Recorded Data.....	7
3.1	Free Field	7
3.2	Office	8
3.3	Cafeteria.....	11
3.4	Road Traffic.....	13
3.5	Downtown Pedestrian Area	14
3.6	Audio-Visual Pilot Recordings.....	14
4	Summary	15

1 Introduction

The present report concerns data collection with DIRAC’s targeted “non-traditional” audio sensor, a six-channel hearing-aid microphone array. The goal of the data collection effort was to record a limited amount of data with this sensor setup to provide input for several WPs in DIRAC. The data recorded is a direct input for the spatial signal processing, scene description and feature extraction work in WP1 and WP2, and through them has indirect relevance for WP4 and WP5. The effort is also relevant for integration (WP6) through its significant input to the development of the application scenario within DIRAC.

As DIRAC’s sensors are non-traditional, there is no publicly available data from these sensors in existence which we could readily use. However, even though existing data is not suitable for the entire system, it can be used for certain parts of the algorithm development. Therefore, it was not desirable (nor planned) to embark on a full-scale database collection effort. Rather, our goal was to record a limit amount of data that can be used in addition to and in conjunction with existing data.

The approach taken was to decompose the acoustic environment into its constituent parts as far as possible. Specifically, we separate the content of a sound source from the properties of the surrounding environment by recording impulse responses from several points in space (e.g., in a room) to the microphones of the microphone array. Convolution of a set of impulse responses with a sound source signal then results in the sound signal as it would have been recorded by the microphone array if the sound source(s) in question had been placed at the specific position(s) in the space. Thereby, existing sound source signals can be used in conjunction with the impulse responses in order to generate a virtual sound field. Being generated digitally, this virtual sound field is a most faithful simulation of the corresponding real sound field since sound propagation in air is known to be linear up to sound pressure levels much higher than used here.

Several non-directional (diffuse) sounds have been recorded with the microphone array since they cannot be simulated as described. Still, the combination of a diffuse sound field with impulse-response simulated directional sources provides a way to combine few recorded diffuse sounds fields with a number of impulse responses and source signals to yield a significant number of acoustic scenes that can be used for algorithm development and training. Finally, a small number of sound scenes with moving sound sources and/or moving microphone array have also been recorded.

This approach has the added advantage of allowing very specific control of acoustic parameters constituting a sound scene. E.g., the ratio of (voice) signal to (background) noise can be specified and varied precisely for algorithm evaluation. Hence, we are able to trigger acoustic parameters in such a way that we obtain benchmark data that will let us evaluate how well algorithms perform under certain situations. The remainder of this report first outlines recording setup and measurement methods and then proceeds with a description of the recording situations and the obtained data.

2 Recording Setups and Measurement Methods

2.1 Artificial Receiver

In the first setup, data was recorded using a head and torso simulator (HATS) on which microphones were mounted. Compared to a human head, the HATS has the advantages of (1) providing inner-ear microphones that record sound pressure near the location of the ear drum and of (2) allowing for a fixed geometry and thereby highly reproducible acoustic parameters of the recordings. Equipment used was the head and torso simulator *Brüel & Kjær Type 4128C* with artificial ears *Brüel & Kjær Type 4158C* (right) and *Type 4159C* (left) plus preamplifiers *Type 2669*. In addition to the inner-ear microphones of the ear simulators sound was concurrently recorded with two 3-channel hearing aid dummies (provided by *Siemens*), one behind each ear, resulting in a total of 8 recording channels. The term hearing aid dummy refers to the microphone array of a hearing aid, housed in its original casing but without any of the integrated amplifiers, speakers or signal processing commonly employed in a hearing aid (cf. Fig. 1).

The recorded analogue signals were pre-amplified using a *G.R.A.S. Power Module Type 12AA*, amplification +20dB (in-ear microphones) and a Siemens custom-made pre-amplifier, amplification +26dB (hearing aid dummies), respectively. A/D conversion used a multi-channel AD/DA-converter (*RME Hammerfall DSP Multiface*) connected to a laptop (*DELL Latitude 610D with 1.7 Ghz Pentium M processor, 1 GB RAM*) via PCMCIA-slot and the digital data was stored either on the internal or an external hard disk.

The software used for the recordings is *MatLab (MathWorks, Version 7.1/7.2 (R14/R2006a))* with a professional tool for multichannel I/O and real-time processing of audio signals (*Soundmex2, HörTech gGmbH*).

For impulse response measurements, sound source sequences were played by an active 2-channel coaxial broadband loudspeaker (*Tannoy 800A LH*). The sequences were generated by scripts developed in-house in *MatLab* and output via the AD/DA-converter.

All data is recorded with a sampling-frequency of 48 KHz and a resolution of 32 Bit. The impulse response measurements obtained from this setup enable us to compose virtual acoustic scenes that are superimposed on the ambient background noise recordings.

2.2 Human Receiver

In the second setup, the HATS was replaced by a real human with the microphone arrays being attached to his ears (cf. Fig. 2). Measurement of impulse responses was not performed with this setup due to the reduced stability of the acoustic parameters from slight head movements. Consequently, no speaker for transmitting signals sequences had to be used and, of course, the in-ear microphone channels are not available in this setup. The 6 remaining channels from the two hearing aid dummies were recorded using an *Edirol FireWire AudioCapture FA-101* AD/DA-converter which is linked to the Laptop (see previous section) via FireWire interface. The *FA-101* audio interface was chosen because it offers improved portability by its feature of power-supply over the FireWire laptop port, which we intend to use in the future. As the presently employed laptop computer does not support this feature, we are currently using an additional custom-made rechargeable regulated power supply to run the AD/DA-converter. The software *Audition 2.0 (Adobe)* was used to do the recordings which were stored on either internal or

Figure 1. Right side of the head and torso simulator with mounted microphone array.

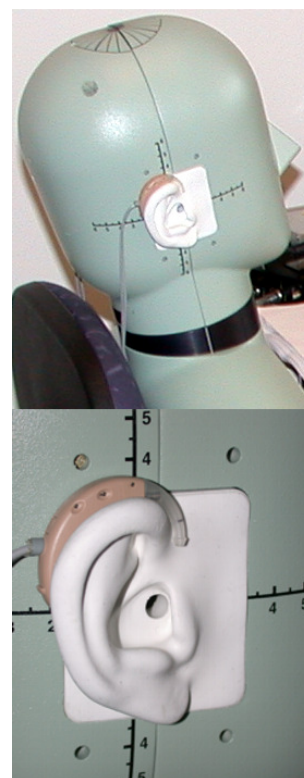


Figure 2. Microphone array attached to human ear.



external hard disk. As this setup is fully portable and battery-powered, it is ideal for the recordings of ambient noise in different situations reported below.

2.3 Impulse Response Measurement Methods

Impulse responses were measured in different situations with two different methods. Impulse response “measurement” is truly an inference process in which a system’s response to a sound signal is fitted to the linear and time-invariant model of the room impulse response. Hence, the estimated impulse response does not only rely on the room but also on the inference process and small non-linear or time-variant properties may show up differently under different estimation methods. The use of two independent measurement methods

therefore enables us to cross-check the consistency of their results while adding only a relatively small amount of time to the automated measurement process.

The first method is based on Maximum Length Sequences (MLS). An MLS is a pseudo random sequence of binary digits which has the spectrum of ideal white noise. By exciting a system with MLS and simultaneously recording its output, the impulse response of the system can be calculated by cross-correlation of the measured response with the original sequence. Here, inverse repeated sequences (IRS) were used, which means that one sequence consists of two MLS, the second one being an inverted copy of the first one. This procedure enhances the immunity to distortion which is important, because for all measurements linear time-invariant systems are assumed, but in reality there are always nonlinear effects which lead to distortions.

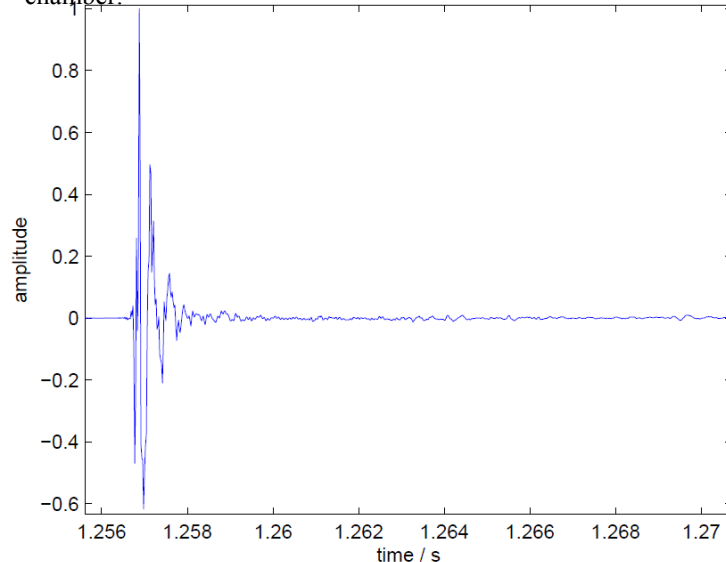
The second method uses logarithmic sine sweeps to calculate the impulse response of a system. After exciting a system with such a signal the recorded response is deconvolved with the original sequence and the impulse response is extracted. A logarithmic sweep requires a shorter time for the measurement in comparison to a linear one and results in advantages regarding to immunity to nonlinearities.

Averaging the system’s response over several repeated sequences reduces the estimated impulse response’s signal-to-noise ratio (SNR). SNR is evaluated by comparing recorded sound signals of test sentences to the convolution of the original sound data with the measured impulse response. For the measurements conducted, SNR reached about 93 dB for the MLS method and about 100 dB for the sine-sweep method.

The measurements were done in an automatic process controlled by *MatLab*. One operation takes about 3 minutes and consists of the following steps:

- impulse response measurement using the MLS-method,
- impulse response measurement using the Sinesweep-method,
- recording of one (anechoic condition) or five (office condition) test sentences played by the loudspeaker.

Figure 3. Example of impulse response recorded in anechoic chamber.



3 Description of Recording Situations and Recorded Data

3.1 Free Field

To simulate a free field situation the measurements were performed in the anechoic chamber of the University of Oldenburg. The HATS was fixed on a computer-controlled turntable (*Brüel & Kjær Type 5960C with Controller Type 5997*) and placed opposite to the speaker in the room (cf. Fig. 4). Impulse responses were measured for distances of 0.8 m and 3 m between speaker and HATS. The larger distance comprises a far-field situation (which is, e.g., commonly assumed by beam-forming algorithms) whereas for the smaller distance some near-field effects may occur. For each distance 4 angles of elevation were set, from -10° to 20° in steps of 10° and for each elevation the azimuth angle of the source to the HATS is varied from 0° to -180° (left turn) in steps of 5° . The recordings are described in a head related coordinate system as depicted in Fig. 5. Hence, a total of 296 ($=37*4*2$) impulse response sets (each set containing 8 individual responses, one for each microphone) and the same number of test sentences were measured and recorded, respectively.

Figure 4. Setup for impulse response measurement in the anechoic chamber.



Figure 5. Coordinate systems for elevation (left) and azimuth (right) angles.

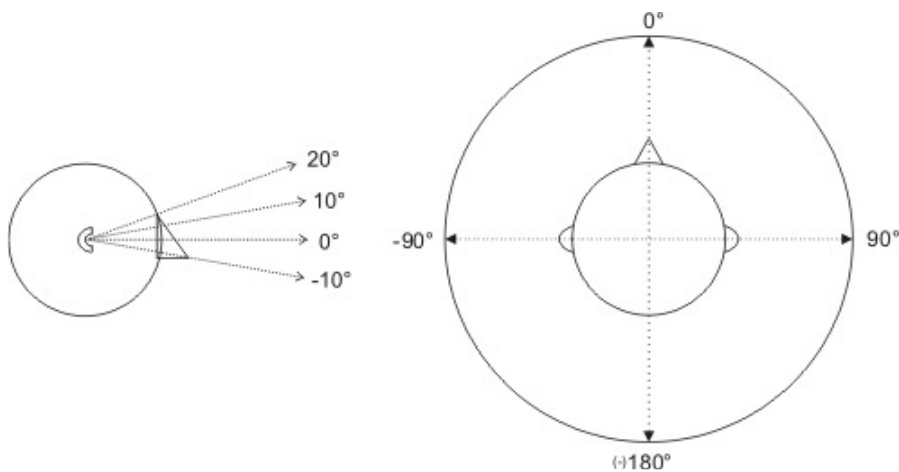


Figure 6. Office at the University of Oldenburg where recordings were performed, showing the head and torso simulator with recording equipment.



3.2 Office

Recordings are done in a typical office room at the University of Oldenburg (cf. Fig. 6). The head and torso simulator was positioned on a chair behind a desk with orientations 0° and 90° , respectively. Impulse responses were measured for four different speaker positions (door, two desks and window) to allow for simulation of sound sources at each of these positions (cf. Fig. 7 and table 1). For measurements at the positions *a* and *d* door and window are opened, otherwise they are closed.

To test the sensitivity of the measurement method to background noise, ventilation installed at the ceiling of the room was switched on for one extra measurement per speaker-position (one head orientation only). In total, this results in 12 impulse responses measured.

Separate recordings of isolated typical office noise sources were performed, consisting of telephone ringing (located in front of the window, 30 seconds recorded for each head orientation) and keyboard typing at the other office desks (positions 1, 2 and 3, respectively, 3 minutes recorded for each head orientation). The ventilation was recorded for 5 minutes (both head orientations). As an extra event the sound of opening and closing the door was recorded 15 times.

Finally, a spatial mix of office noise is recorded with a human receiver (instead of the HATS) for 15 minutes, covering typical sounds such as typing, telephone-ringing, a person walking and noise from the window.

Table 1. Summary of locations of impulse response measurements (a, b, c, d) and of noise sources (1,2 3, telephone, ventilation) that were measured for each of two possible head positions (I, II).











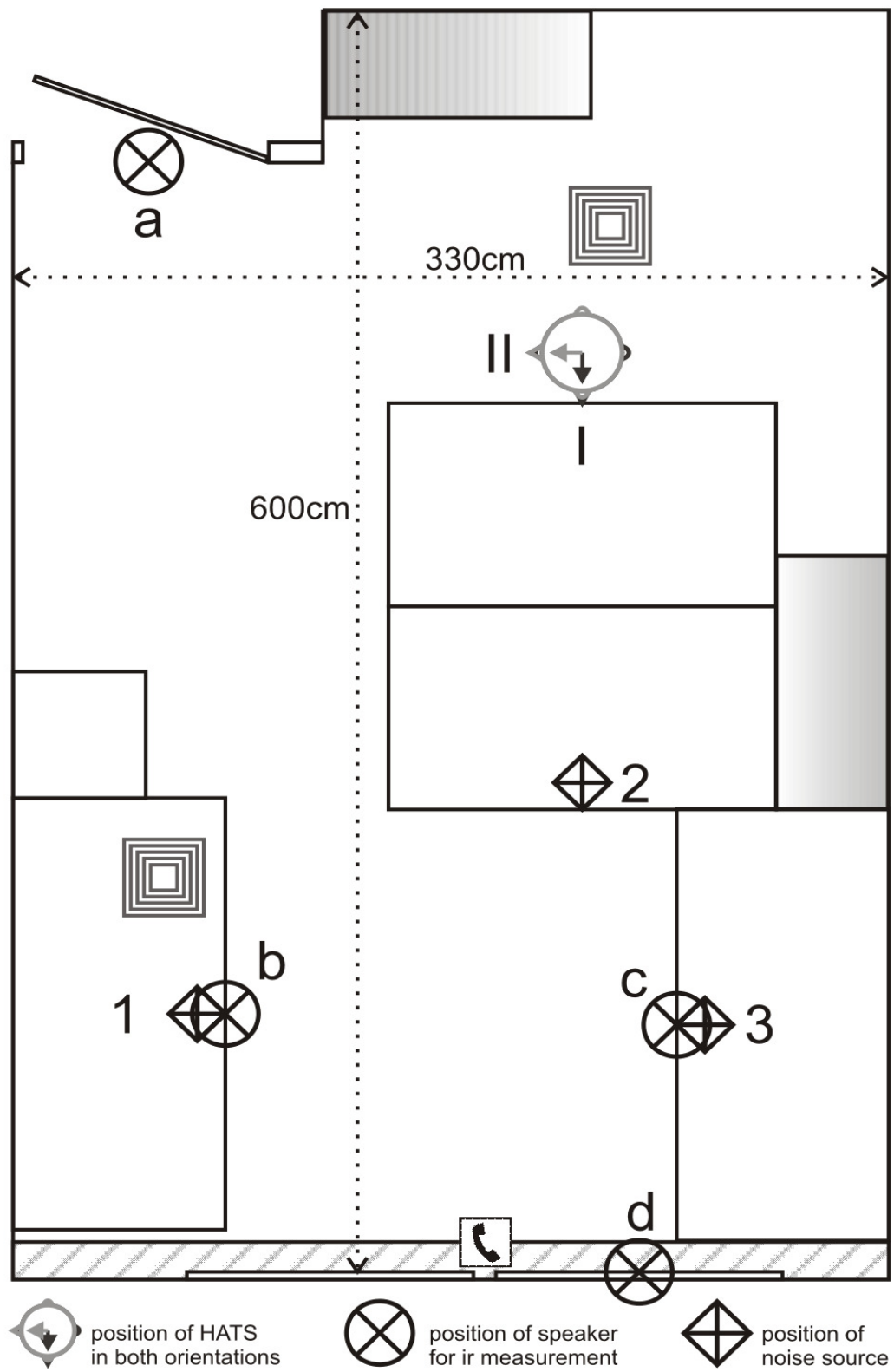
orientation of HATS			total amount
impulse response	 a b c d	 a b c d	12 impulse responses
keyboard typing	 1 2 3	 1 2 3	6 min
telephone ringing			1 min
ventilation			10 min
door	•	•	approx. 2 min

Figure 7. Schematic drawing of the office environment with microphone locations.



3.3 Cafeteria

15 minutes of ambient noise were recorded with the mobile (human receiver) setup in the cafeteria of the natural sciences campus of the University of Oldenburg. The recording was performed at lunchtime when the cafeteria was fully occupied. Ambient noise consisted mainly of unintelligible babbling from simultaneous conversations, occasional parts of intelligible speech from nearby speakers, clanking of dishes and chairs scratching on the stone floor (see Fig. 13, left panel, for an example).

In a second activity the artificial setup was used to measure impulse responses from different positions (see Fig. 8) and collect additional noise from the cafeteria. Again the busy lunchtime was exploited to have realistic conditions (cf. Fig 9).

Figure 8. Schematic drawing of the cafeteria environment with locations of impulse response measurements (A, B, C, D, E, F) for both of the preset orientations of the head (1, 2).

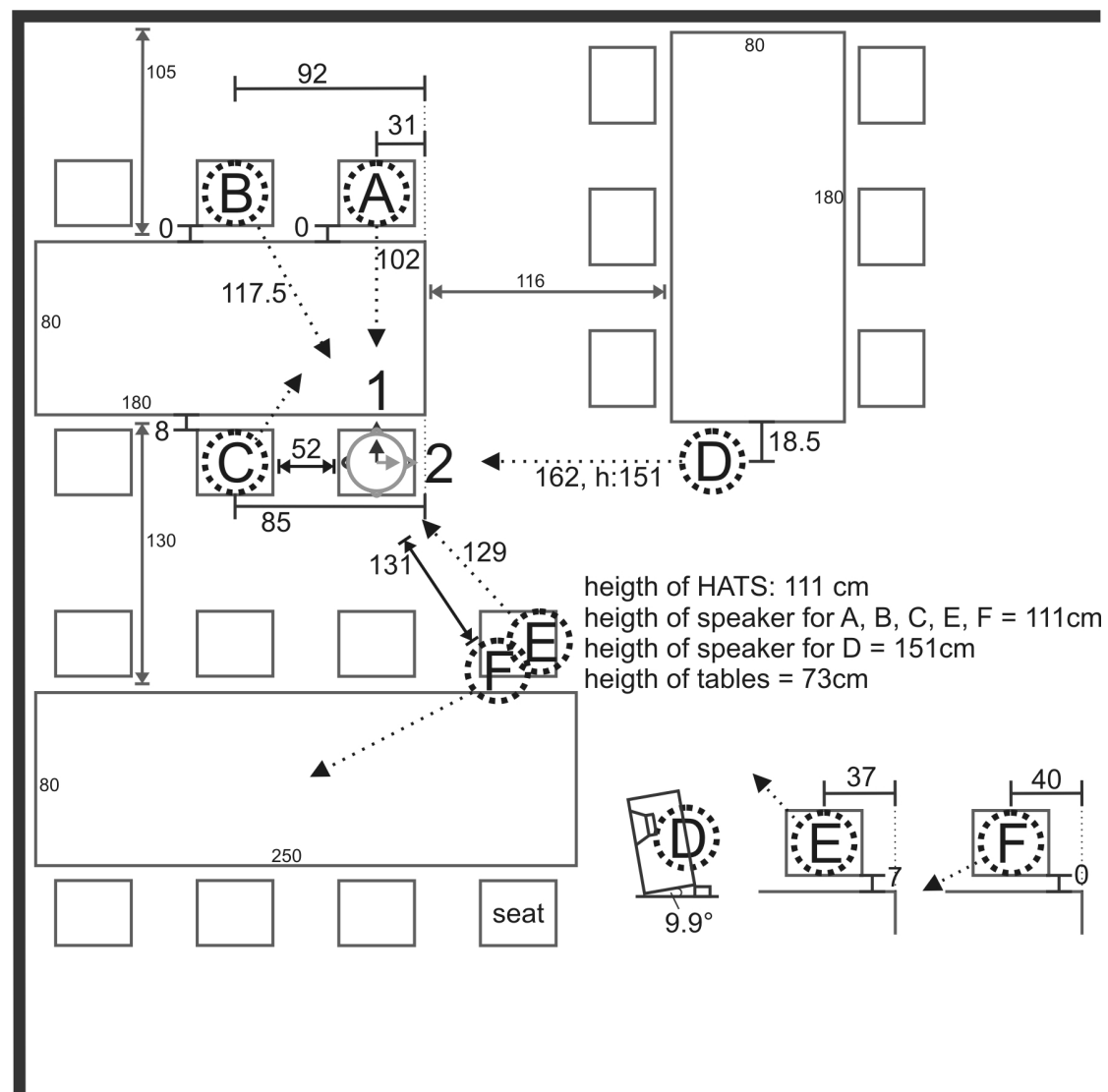


Figure 9. Cafeteria at the University of Oldenburg where recordings and measurements were conducted.



3.4 Courtyard of University

Measurements in the courtyard of the natural sciences campus of the University of Oldenburg were conducted analogous to the “office” and “cafeteria” recordings described above (cf. Fig. 10 and Fig. 11).

A bicycle- and a footpath take course through this yard, so people passed by mainly producing babbling noise, sounds from stepping and mechanical sounds from bicycles containing sudden events like ringing and squeaking of brakes. Continuous noise from trees and birds in the surrounding was also present.

Figure 10. Schematic drawing of the courtyard of the university with locations of impulse response measurements (A, B, C, D, E, F) for both orientations of the head (1, 2).

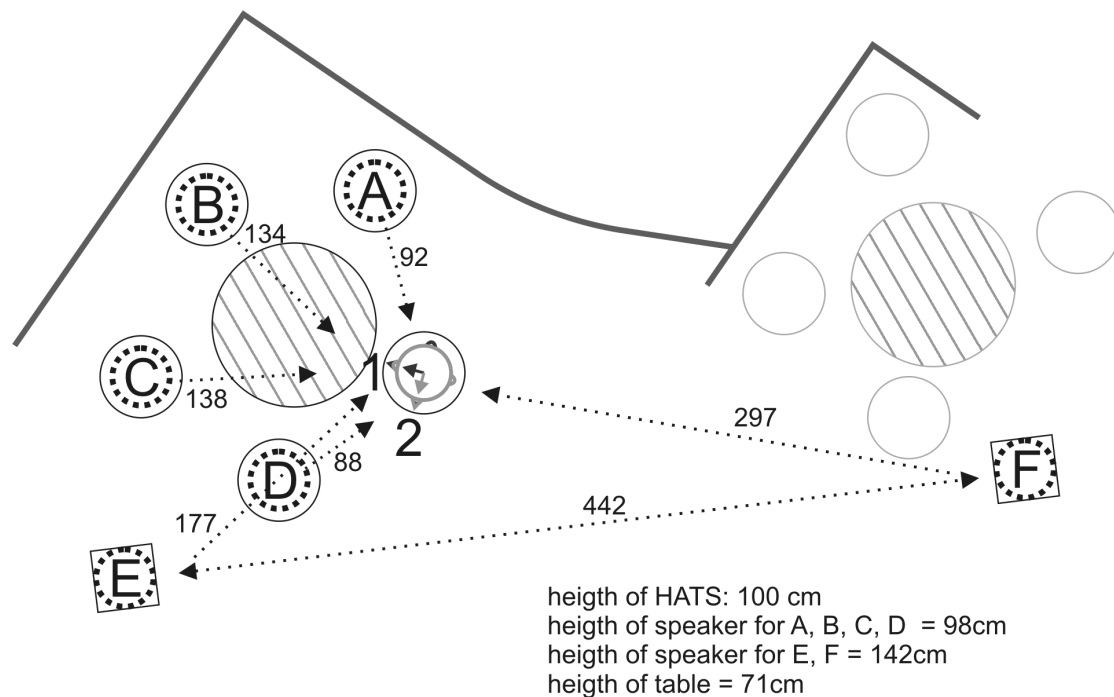


Figure 11. Courtyard of the natural sciences campus of the University of Oldenburg



3.5 Road Traffic

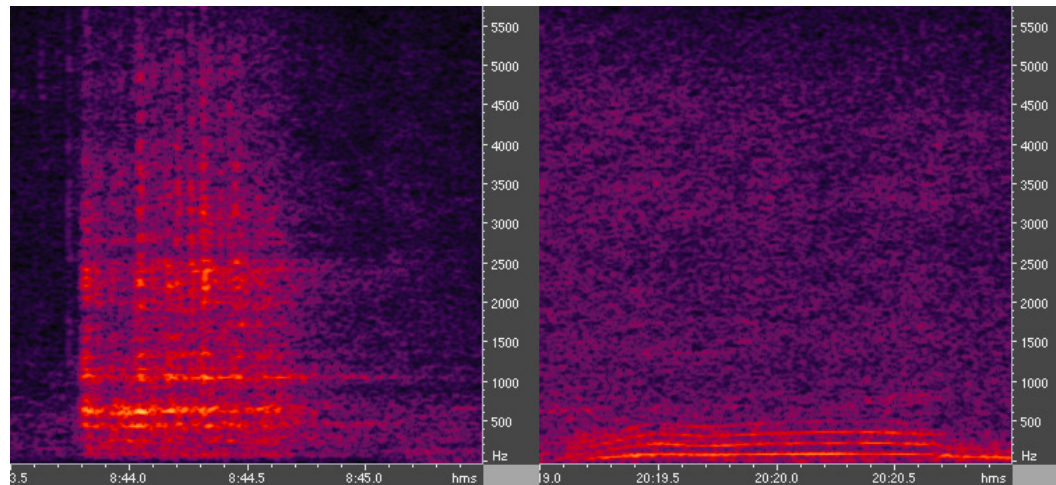
Using the mobile (human receiver) setup, traffic noise was recorded near a heavily used crossroad (cf. Fig. 12). Two recordings with a duration of 5 minutes each were obtained, first with a distance of about 3 meters to the road and afterwards very close-by to simulate the situation of a person who wants to cross a road. During recording, certain events in the traffic scene were noted, e.g., green phases of traffic lights, trucks passing by and squeaking of breaks or tires. Motor vehicles passing by and starting at green traffic lights accounted for the main portion of ambient noise (cf. Fig. 13, right panel for an example).

Figure 12. Intersection where traffic noise recording was carried out.



Figure 13. Example spectrograms of typical events in recorded scenes: chair scratching on the floor of the cafeteria (left panel) and accelerating motorbike at a green traffic light (right panel). Abscissa

represents time, ordinate frequency, light colours high spectro-temporal signal power, dark colours low power.



3.6 Downtown Pedestrian Area

This recording was conducted using the mobile setup on a Saturday around noon time in the city centre of Oldenburg with many pedestrians in the area as seen in Fig. 14. The whole recording covers about one hour and contains a round tour through the pedestrian area with two stops of about five minutes each for recordings from fixed positions. Ambient noise mainly consists of pieces of conversation and babbling in the background, music from stores and from outside presentations.

Figure 14. City centre of Oldenburg where recordings were performed.



3.7 Audio-Visual Pilot Recordings

At the DIRAC project meeting in Leuven (September 5/6, 2006) audio-visual test recordings were done in cooperation with the research group from the Center for Machine Perception at

the Czech Technical University (CTU) in Prague. Together with the mobile visual recording setup from CTU the recordings were performed in an office room at the building of the Katholieke Universiteit Leuven and outside in a parking lot. Different setups of the visual equipment were tested, e.g., the frame rate for image acquisition was varied and pictures were taken in colour and black and white. Overall six audio-visual recordings with a total duration of about seven minutes were captured.

At the same venue, about 80 minutes of audio-only recordings were performed, capturing a meeting in a large conference room using the HATS with 6-channel hearing aid microphones (no inner ear microphone channels).

4 Summary

The sound data acquisition effort carried out has been described in some detail in this report, outlining motivation, methods and data obtained. In total, the recordings comprise 332 impulse responses, 303 min. audio recordings and 7 min. pilot audio-visual data recordings, captured from six situation setups, see table 2. The data collected is already starting to be used for the work in WP1 and WP2. During the course of the recording, our six-channel body-mounted microphone array has been integrated into a recording setup that is fully portable and mains independent. This setup forms the basis for the audio part of DIRAC's application scenario audio-visual cognitive aid. Once an audio-visual setup is ready, we anticipate to be able to move smoothly to the new hardware platform.

Table 2. Summary of all measurements and recordings.

	<i>Data recorded</i>
Free field	296 different impulse response sets
Office	296 test sentences, total duration 55 min. 12 impulse response sets
Cafeteria	40 test sentences, total duration 8 min. 13 recordings of ambient noise, total duration 34 min. 12 impulse response sets
Courtyard of university	2 recordings of ambient noise (HATS), total duration 14 min. 1 recording of ambient noise (human setup), total duration 15 min. 12 impulse response sets
Road traffic	12 test sentences, total duration 3 min.
Downtown	1 recording of ambient noise, total duration 24 min. 2 recordings of ambient noise, total duration 10 min.
Pilot a/v recordings	3 recordings of ambient noise, total duration 60 min. 6 recordings of a/v scenes, total duration 7 min.
Total	1 meeting audio recording, total duration 80 min. 332 impulse responses 303 min. audio recordings 7 min. audio-visual recordings