A Multimodal-Corpus Data Collection System for Cognitive Acoustic Scene Analysis

Julius Georgiou*, Philippe Pouliquen*+, Andrew Cassidy*+, Guillaume Garreau*, Charalambos Andreou*, Guillermo Stuarts*, Cyrille d'Urbal*, Andreas G. Andreou*+
*Department of Electrical and Computer Engineering, University of Cyprus, Nicosia 1678, Cyprus
+Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA
Email: julio@ucy.ac.cy, {andreou,pouliquen,cassidy}@jhu.edu

Susan Denham, Thomas Wennekers, Robert Mill
Department of Psychology
and the Centre for Robotics and Neural Systems,
University of Plymouth, United Kingdom
Email: S.Denham@plymouth.ac.uk

István Winkler, Tamás Bőhm, Orsolya Szalárdy
Institute for Psychology, Hungarian Academy of Sciences
Budapest, Hungary
Email: iwinkler@cogpsyphy.hu

Georg M. Klump, Simon Jones
Department of Biology and Environmental Sciences
University of Oldenburg, Oldenburg, Germany
Email: georg.klump@uni-oldenburg.de

Alexandra Bendixen
Institute of Psychology, University of Leipzig
Leipzig, Germany
Email: bendixen@cogpsyphy.hu

I. Abstract

We report on the design and the collection of a multi-modal data corpus for cognitive acoustic scene analysis. Sounds are generated by stationary and moving sources (people), that is by omni-directional speakers mounted on people's heads. One or two subjects walk along predetermined systematic and random paths, in synchrony and out of sync. Sound is captured in multiple microphone systems, including a four MEMS microphone directional array, two electret microphones situated in the ears of a stuffed gerbil head, and a Head Acoustics, head-shoulder unit with ICP microphones. Three micro-Doppler units operating at different frequencies were employed to capture gait and the articulatory signatures as well as location of the people in the scene. Three ground vibration sensors were recording the footsteps of the walking people. A 3D MESA camera as well as a web-cam provided 2D and 3D visual data for system calibration and ground truth. Data were collected in three environments ranging from a well controlled environment (anechoic chamber), an indoor environment (large classroom) and the natural environment of an outside courtyard. A software tool has been developed for the browsing and visualization of the data.

II. Introduction

How much can we learn about what is going on in the world around us simply by listening? What can we tell about the behavior or presence of other living creatures in our environment using only information derived from sounds? And how does the sound generating interaction of animate entities with the plethora of inanimate objects gives clues about context? Can we tell whats in the mind of living creatures only by their sound generating actions and interactions with the environment? How does one link in a coherent and systematic way the knowledge from physics/acoustics, human auditory neuroscience and psychophysics together with engineering and computer science to engineer computational models and algorithms for robust sound understanding and acoustic scene analysis. And, if our methodology relies on engineered systems that continuous learn and adapt to the environment what additional challenges algorithmic and computational do we need to address?

The multi-modal data corpus collection discussed in this paper is aimed at providing the data to support the science and engineering aimed at answering questions such as the above with the end goal of engineering real-time system which uses the sounds generated by other creatures as well as changes in the sounds which bounce off them to decide whether there is a living being nearby, where it is and what it is doing. In building this system we draw on biological inspiration in the design and optimization of a computational processing architecture and sensors, also emulating the perceptual processing strategies employed by human and animals. Our scientific understanding builds on many years of work by ourselves and others, as well as experiments which designed to answer specific scientific questions; questions related to how we interpret the incoming signals in terms of objects in the environment. How do we decide whether this newly observed movement or sound was generated by someone we already know about or someone new, and how can we make these decisions adaptively so we can interact in a meaningful way with the world? Finally how

978-1-4244-9848-2/11$26.00©2011 IEEE

can we build representations of normal signals so that we can recognize when things change unexpectedly?

III. Science and Engineering

The collected multi-modal data corpus will be employed to advance the state of the art in auditory perception, auditory neurophysiology and modeling as well as the neuromorphic cognition engineered of real-time systems in the European FP7 project SCANDLE [1]

From a science perspective, the sound protocols were selected with the following two objectives to facilitate perceptual experiments in humans. These experiments are designed to further elucidate and understand the mechanisms of (i) integration between multiple coherent cues employed by the brain to separate sound sources and form stable auditory object representations and (ii) explore the conditions for the detection and formation of a new perceptual object by assessing the role of the old+new heuristic in human perception. Complementary to the perceptual experiments, the data corpus will be employed in neurophysiological experiments in an animal model, the Mongolian gerbils (Meriones unguiculatus), thus combining animal psychophysics and recording from cortical neurons to investigate the neural mechanisms underlying the MisMatch Negativity (MMN) response. It is assumed that mismatch negativity to be important for parsing incoming sounds into representations of discrete sound sources, or auditory objects.

The data collected will also help formulate a model of auditory processing consistent with connectivity in the auditory system, and capable of autonomous perceptual organization as well as investigate the ability of the model to simulate typical phenomena of autonomous flexible auditory perceptual organization. More specifically it will aid in the formulation of a neuro-computational model of auditory processing in a form compatible with neuromorphic hardware and capable of asynchronous, stimulus-driven, real-time classification of passively detected sounds; i.e. those emitted by entities in the environment. This model will be extended to include recurrent connections with higher cortical areas, in order to investigate the parsing of incoming sound streams through the formation of expectations and the emergent representations of auditory objects. The active acoustic data collected with the micro-Doppler sensors will be employed to investigate extensions to the auditory processing model to allow classification of spectrotemporal patterns in sounds resulting from active exploration of the environment (active audition). Finally we will explore ways of sonifying spiking neuron network activity in order to reveal cognitively relevant states and properties of self-organized activity to an observer, e.g., the presence of several objects, their relative salience in the streaming paradigm, bi-stable phenomena, active memory, decisions and formulate extensions to the models to integrate and form composite representations of passive and active sources of acoustic information.

From an engineering perspective, our aim is a distributed cognitive acoustic scene analysis system. The integration and deployment of such cognitive sensor network system that employs both passive and active acoustic sensors relies on a thorough understanding of the limitations in the individual components. The data collected will help us assess the limits on detection and classification of movement, localization of objects, as well as the ability of the wireless communication system to support such distributed processing. For example to deal with the wide dynamic range of signals in the natural environments, we will investigate neural mechanisms for automatic gain control of the acoustic front ends and cochlear models to compensate for the enormous distance-dependent variability of sonar returns and passively detected sounds. The collected data corpora for passive and active acoustic signatures in different situation environments will also be important in learning internal models of the natural environment necessary for robust system performance. Once the system is functional, the data corpus will be employed to assess its performance and limitations of the architecture on a well defined set of problems with animate and inanimate articulated moving entities. Having a standard data set is also important when comparing the performance of biological systems to engineered systems.

As such we aim at optimizing the design of bio-inspired computing architectures, and solving problems associated with massively parallel computation and sensory information processing. Here too we draw on biology to understand the constraints and costs associated with design choices in the brain. The key idea here is that brains have evolved to solve problems in natural environments that constrain the problem. For example, living things exist in a three dimensional world and laws of physics determine the sounds that are produced when their motor system actions result in interactions with objects in the environment. For example, when moving articulated objects produce interact with sound waves, because of the Doppler effect there is a frequency shift in the observed frequency of the acoustic waves that gives out information on the dynamical structure of the moving object as well as its context with in the environment. Similarly, a car engine produces periodic sounds that have spectral content that is unique to the physical components on the car. At the cognitive level a human will often say “it sounds like a car”, in this instance car is a projection of the physics of sound through the acoustic signal and onto a symbolic representation, the class of objects defined by the word “car”.

Our work is ultimately aimed at a methodology for determining the architecture of specialized embedded systems and for compiling large scale brain models for real-time operation in situated environments. In the coming years we expect this technology to open up many interesting and valuable application areas in the realm of truly Neuromorphic Cognitive Machines. By using only sound, issues of the invasion of privacy are reduced, and so engineered systems could be used to intelligently monitor home environments so as to facilitate independent living for the elderly. Interactive gaming systems with a better representation of peoples’ behavior, remote monitoring of animals or fish for automating identification, and detecting life where visual contact is obscured.

978-1-4244-9848-2/11$26.00©2011 IEEE

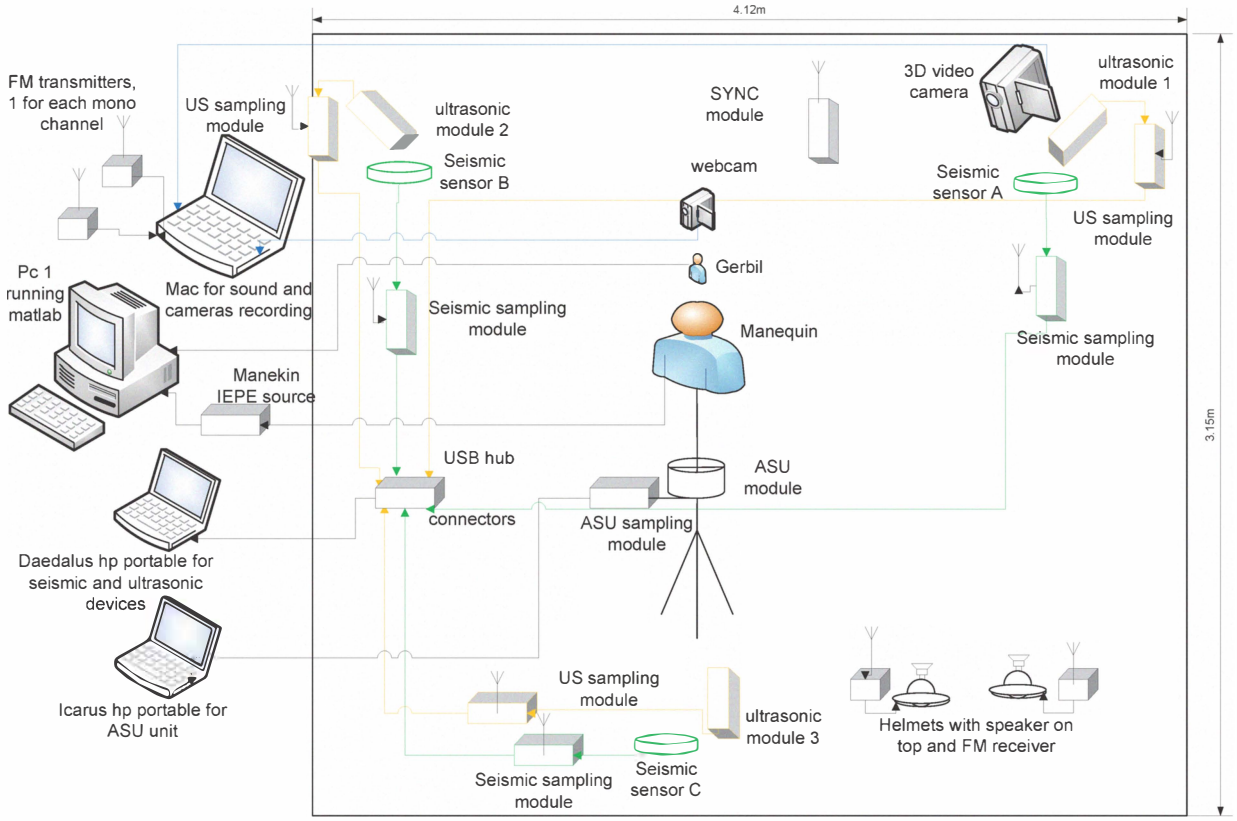

Fig. 1. Data acquisition system architecture. The dimensions and location of individual components correspond to the approximate location of the equipment in an indoor environment (large classroom)

IV. System Architecture

A system-level architecture or the data collection instrumentation is shown in Fig. 1. Three computers with commercial off the shelf data acquisition cards from National Instruments [2] [3] were employed for the data acquisition. The data acquisition cards were controlled via low-level C-software routines and at the higher level through Matlab (see Fig. 5). Synchronized start and stop in the data acquisition was provided through the computer serial ports and a wired connection. All data were directly stored on a Network Attached Storage RAID system (QNAP TS-639 Pro) and with a backup and the end of the data collection on two separate USB drives.

A. Sensors

Sound is captured through multiple systems including a four MEMS [4] microphones directional array [5] [6] [7], two high quality Knowles electret microphones [8] situated in the ears of a stuffed Gerbil head (see Fig. 2). The data for human perceptual experiments were recorded using a Head Acoustics [9] head-shoulder unit with built-in ICP microphones. Three micro-Doppler units [10] operating at different frequencies were employed to capture the articulatory signatures and location of the people in the scene. Three ground vibration sensors each with a vector output (x-y-z components) [11] were employed to record people footsteps. Video recorded with a MESA 3D camera [12] and a web camera provided visual information and ground truth for the location of the people and their body movements.

B. Mobile sound generation system

The sounds were generated using Matlab and played back using the stereo audio output of a 17 inch MacBook Pro computer (version Fall 2009) running OSX Leopard 10.5.6. The stereo line output of the computer was split into two mono-aural signals using a custom cable. These signals were fed in the transmitter microphone input of an Audio-Technica Pro 88W/T wireless microphone [13]. Two such transmitter units were employed together with their corresponding receiver units to transmit the audio signal from the left and right channels of the computer to the mobile speaker units. The two wireless microphones were tuned to different frequencies and hence there was no interference.

The mobile speaker units consisted of a Gallo Acoustics A'Diva Ti omnidirectional speaker [14] (Fig. 3) mounted on a helmet and driven by a custom designed amplifier. The frequency response of the speaker was measured by HTLabs [15] and is shown in Fig. 4. The input of the

978-1-4244-9848-2/11$26.00©2011 IEEE

Fig. 2. Stuffed Gerbil with microphones in the ears

Fig. 3. Omni-directional Gallo Diva TI speaker used in the experiments

amplifier was connected to the output of the Audio-Technica Pro 88W/T wireless microphone receivers. The amplifier and wireless receiver units as well as batteries were carried by the subject in backpacks. Two such units were available allowing two individual subjects to produce independent sounds while walking and behaving in the test environment. The helmet mounted speakers can be seen in the bottom of Fig. 6.

C. Data Collection Environments

Data were collected in three environments: an anechoic chamber, a classroom (Fig. 1) and in an outside backyard (Fig. 6).

V. Sound Protocols

Three sets of sounds were used. The first one was delivered using a streaming protocol [16], [17] *A*BABABA ..... For this part, three different sets of sounds were generated, one with pure tones, the second with resolved harmonic complexes

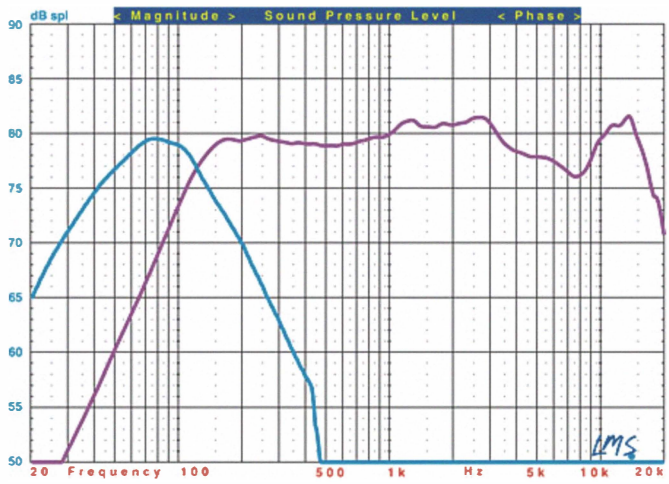

Fig. 4. Frequency response of the Gallo A'Diva TI speaker used in the experiments (purple line). Measurements done at HTLabs

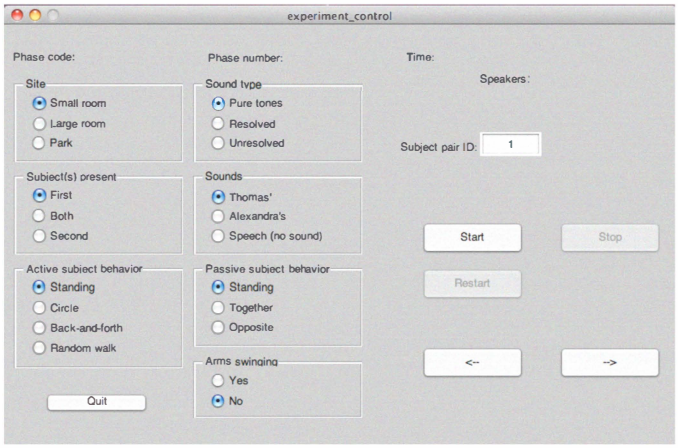

Fig. 5. Experiment control panel (Matlab GUI)

Fig. 6. Photograph of the experimental setup in the outside court-yard. Two subjects are shown wearing the helmets with speakers and the backpacks that have the wireless sound generation equipment .

978-1-4244-9848-2/11$26.00©2011 IEEE

and the third consisting of unresolved harmonic complexes. The sounds were emitted by a single speaker on the helmet of a person participating in the experiment or alternating between the speakers on two speakers situated on the helmets of the two people participating in the experiment. The single person or two people were asked to move in a prescribed area while the sound was emitted from their helmets. The experiments were conducted in an area defined by a circle. The approximate diameter of the circle was 3m for the indoors control environment (anechoic chamber), 6m for the indoor large classroom and 10m for the outdoors courtyard experiment. The systematic movement patterns were defined as (1) walking on the circumference of a circle, (2) walking radially from the periphery towards the center and back along the radius of the circle and (3) walking on a random trajectory. The first set of sounds were selected because the required binding of the different frequency components.

The second set of sounds was designed to vary along three different dimensions, (i) sound complexity (pure tones to complex tones), (ii) sound dynamics (speed of frequency modulation) and (iii) sound source location and movement. These sounds were delivered by a single loudspeaker mounted on a participants head. The participant either walked on the perimeter of a circle or stood still at a given location of the scene during the sound presentation. These sounds were selected because they could provide a rich environment for testing more sophisticated models of sound source discrimination.

The third set of sounds was natural speech. Two participants walked together about the room on a random trajectory and conducted a non-scripted conversation with each other. Speech was included to provide a fully natural scenario for testing models.

VI. Graphical User Interface for Data Visualization

To make the data analysis easier, a matlab Graphical User Interface (GUI) (Fig. 7) was developed that allows user to have access to all the data recorded in the multimodal data corpus, including the 2D and 3D camera video. From a graphic interface, any configuration and any data can be selected and then displayed. In addition to the actual GUI interface, a Matlab toolbox and a programmer's guide was also developed to facilitate extraction and analysis of the data from the corpus.

VII. Discussion and Conclusions

A major challenge in the design of the system to collect the corpus was the speaker. The broadband sounds necessitate a speaker with flat response from a approximately 100Hz to 20Khz. In a preliminary set of experiments, a high quality tweeter transducer [18] that has been previously used for acoustic experiments deemed problematic as a resonance at about 500Hz distorted the stimuli.

By listening the recorded sounds using stereo headphones, it appears that the localization cues are recorded properly: one can perceive the sound source moving around. The recorded

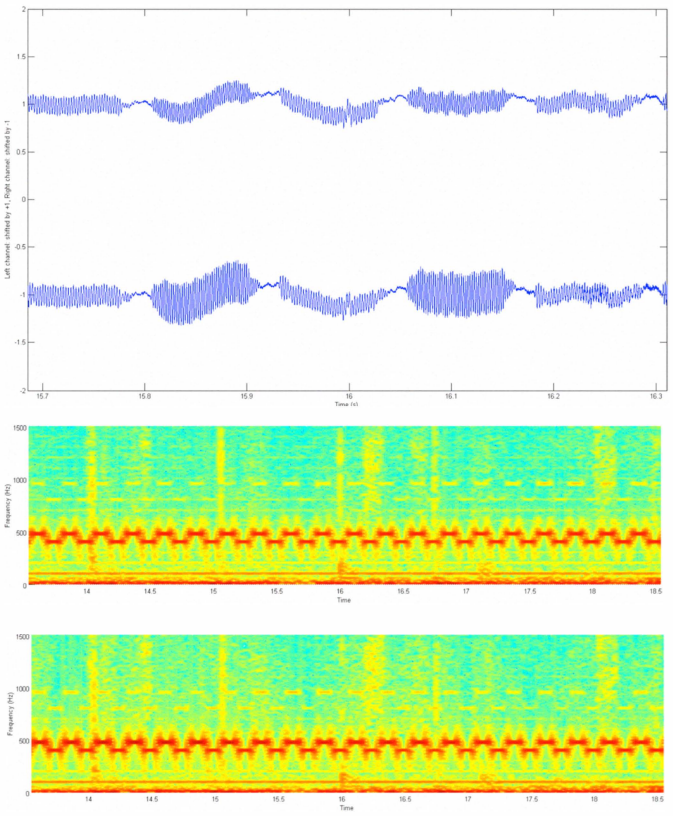

Fig. 8. Experimentally recorded data from from the two ICP microphones in the HeadAcoustics mannequin head. The audio data from one experimental phase was loaded and displayed by Matlab functions *bin2wav*. The data displayed here correspond to an experimental condition where alternating high and low pure tones were played, both from the same speaker and the subject was moving in a circle. The experiment was done in in the anechoic chamber (indoors control environment). The binary data was read in Matlab and it was re-sampled at 40 kHz, normalized in amplitude and subtract the mean of each channel to remove the dc offset. The waveforms (top) and spectrograms (bottom) of roughly the same short time-interval in two channels of the ICP microphones are plotted. Note that the time axes of the time domain waveforms are not the same as the time axes of the spectrograms. The vertical striations on the spectrogram probably correspond to footsteps and other noises made by the subject.

signals have quite a large low frequency energy component due to the power lines that can be eliminated by high-pass filtering. On all the recordings, one can see a very weak harmonic of each stimulus tone, supposedly caused by the slightly imperfect reproduction of the sinusoids by the sound playback system; this is not a significant problem.

In this paper we presented the details and information on a data collection system employed to acquire a multimodal corpus for data corpus for cognitive acoustic scene analysis. The corpus consists of data collected in three sites, a well controlled environment (anechoic chamber), an indoor environment (large classroom) and the natural environment of an outside courtyard. We reported on the experimental protocols for the data collection, the experimental setup, software tools for data conducting the experiments as well as a MATLAB graphical user interface (GUI) for browsing the multi-modal database. The total size of the corpus is approximately 500

978-1-4244-9848-2/11$26.00©2011 IEEE

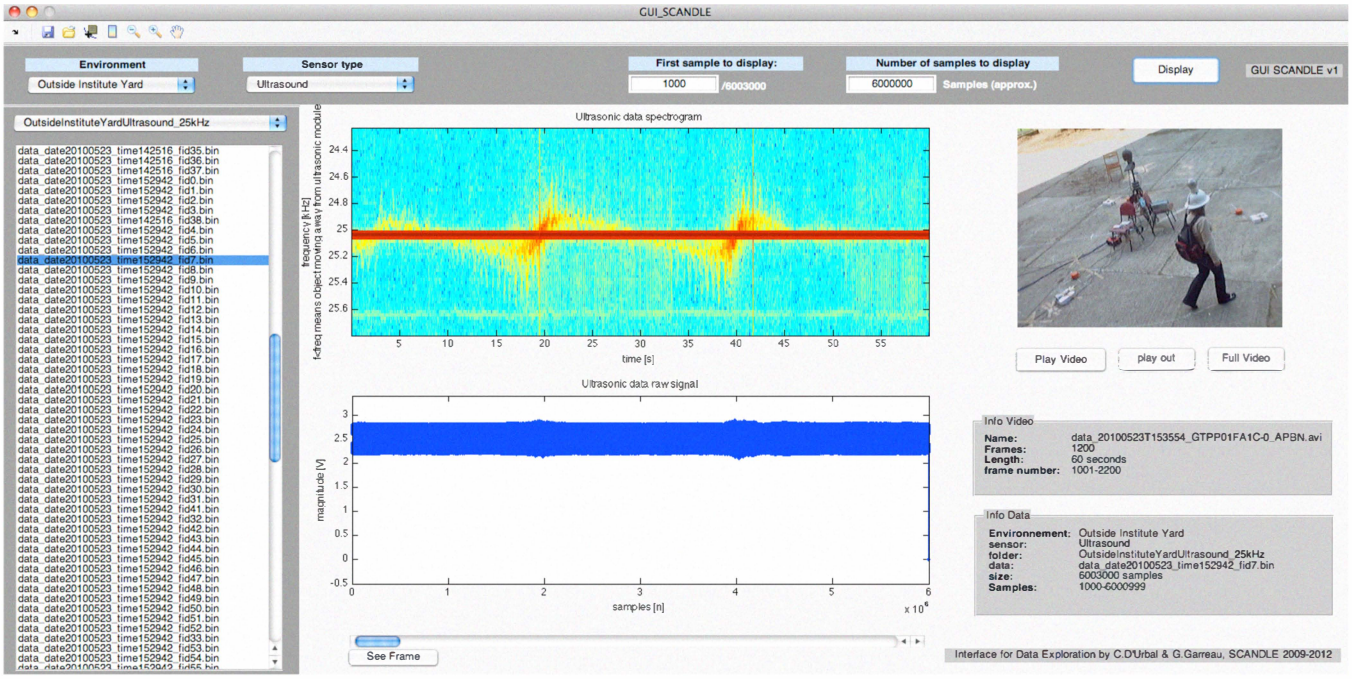

Fig. 7. Graphical user interface for browsing the multi-modal data corpus. The GUI interface displays a signal window in the middle with data from one of the micro-Doppler sensors. In the right side of the GUI one can see the video frame with the subject performing the required task -outside environment-.

Giga-bytes.The data acquisition system employed in to acquire this data corpus is a unique system [19] that allows synchronization of data capture with the unprecedented 5uS resolution! The synchronization of all computers in the system is done through dedicated serial RS232 lines and low-level Matlab function calls. The multimodal data corpus will become available at *http://www.scandle.eu/* in the Spring of 2012.

ACKNOWLEDGMENT

This work was supported by the European Research area Specific Targeted Project SCANDLE (IS-FP7-231168). We thank for the assistance of Zsuzsanna D'Albini and Judit Roschéné Farkas during the experiments. We are grateful to all the staff at the Institute for Psychology, Hungarian Academy of Sciences that supported the data collection team for the duration of the experiments.

REFERENCES

[1] EU-FP7-Project, "SCANDLE: acoustic SCene ANalysis for Detecting Living Entities," *http://www.scandle.eu/*, Mar 2011.

[2] NI, "24-bit, 204.8 kS/s dynamic signal acquisition and generation," *NI-4662 data sheet*, Apr 2006.

[3] ——, "R series multifunction RIO with Virtex-5 LX30 FPGA," *Datasheet*, vol. PCIe-7841R, Feb 2010.

[4] Knowles, "Spm0408he5hmicrophone," *Datasheet*, pp. 1–10, Sep 2009.

[5] G. Cauwenberghs, A. G. Andreou, J. West, M. Stanacevic, A. Celik, P. Julian, T. Teixeira, C. Diehl, and L. Riddle, "A miniature, low-power , intelligent sensor node for persistent acoustic surveillance," *SPIE Proceedings, Unattended Ground Sensor Technologies and Applications VII Conference*, Apr 2005.

[6] P. Julian, A. G. Andreou, L. Riddle, S. Shamma, D. Goldberg, and G. Cauwenberghs, "A comparative study of sound localization algorithms for energy aware sensor network nodes," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 4, pp. 640– 648, Jul 2004.

[7] P. Julian, A. G. Andreou, G. Cauwenberghs, M. Stanacevic, H. Goldberg, P. Mandolesi, L. Riddle, and S. Shamma, "Field test results for low power bearing estimator sensor nodes," *2005 IEEE International Symposium on Circuits and Systems (ISCAS 2005)*, vol. 5, pp. 4205 – 4208, Apr 2005.

[8] Knowles, "Knowles FC-23329-P07," *Datasheet*, Nov 2005.

[9] Head-Acoustics, "HSU IIL2 head shoulder with ICP microphones," *Datasheet*, Sep 2008.

[10] Z. Zhang, P. Pouliquen, A. Waxman, and A. G. Andreou, "Acoustic micro-Doppler gait signatures of humans and animals," *41st Annual Conference on Information Sciences and Systems (CISS 2007)*, pp. 627–630, 2007.

[11] Mark-Products, "Mark products 4.5Hz 3 component L15B sensor," *www.seismicnet.com/geophone/index.html*, Feb 2011.

[12] MESA, "3d time of flight camera," *http://www.mesa-imaging.ch/index.php*, Feb 2011.

[13] Audio-Techninca, "PRO-88W VHF wireless lavalier microphone system," *www.audio-technica.com*, Feb 2009.

[14] Gallo-Acoustics, "Adiva Ti Owner's manual," *www.roundsound.com/adiva-ti-speakers.htm*, Sep 2008.

[15] S. Guttenberg, "Anthony Gallo acoustics A'Diva Ti speaker," *www.hometheater.com/content/anthony-gallo-acoustics-adiva-ti-speaker-system*, Apr 2010.

[16] L. P. van Noorden, "Temporal coherence in the perception of tone sequences," *Ph.D. Dissertation*, Feb 1975.

[17] A. S. Bregman, "Auditory Scene Analysis," *The MIT Press*, May 1990.

[18] Vifa, "XT/DX 1" tweeter," vol. XT225tg30-04, Feb 2009.

[19] P. O. Pouliquen, A. Cassidy, G. Garreau, J. Georgiou, and A. G. Andreou, "A wireless architecture fro distributed sensing/actuation and preprocessing with microsecond synchronization," *45th Annual Conference on Information Sciences and Systems (CISS 2011)*, Mar 2011.

978-1-4244-9848-2/11$26.00©2011 IEEE