Heterogeneous Sensor Database in Support of Human Behaviour Analysis in Unrestricted Environments: The Audio Part

Stavros Ntalampiras†, Todor Ganchev†, Ilyas Potamitis‡, Nikos Fakotakis†

†University of Patras, Department of Electrical and Computer Engineering
Panepistimioupoli, 26500, Rio, Patras, Greece
E-mail: sntalampiras@upatras.gr, tganchev@wcl.ee.upatras.gr, fakotaki@upatras.gr

‡Technological Educational Institute of Crete, Department of Music Technology and Acoustics
Daskaliki-Perivolia, 74100, Rethymno, Crete, Greece
E-mail: potamitis@wcl.ee.upatras.gr

Abstract

In the present paper we report on a recent effort that resulted in the establishment of a unique multimodal database, referred to as the PROMETHEUS database. This database was created in support of research and development activities, performed within the European Commission FP7 PROMETHEUS project, aiming at the creation of a framework for monitoring and interpretation of human behaviours in unrestricted indoors and outdoors environments. In the present paper we discuss the design and the implementation of the audio part of the database and offer statistical information about the audio content. Specifically, it contains single-person and multi-person scenarios, but also covers scenarios with interactions between groups of people. The database design was conceived with extended support of research and development activities devoted to detection of typical and atypical events, emergency and crisis situations, which assist for achieving situational awareness and more reliable interpretation of the context in which humans behave. The PROMETHEUS database allows for embracing a wide range of real-world applications, including smart-home and human-robot interaction interfaces, indoors/outdoors public areas surveillance, airport terminals or city park supervision, etc. A major portion of the PROMETHEUS database will be made publically available by the end of year 2010.

1. Introduction

Prediction and interpretation of human behaviour using probabilistic structures and heterogeneous sensors (PROMETHEUS) is a project[1] funded under the umbrella of EC FP7. PROMETHEUS project promotes research on probabilistic inference algorithms within the paradigm of recursive Bayesian estimation to the problem of online tracking of multiple people in a scene, and the identification of the interaction amongst them and with the environment. The core research components of the project are in the representation of uncertainties arising from multiple modalities including visual, acoustical and infrared, the fusion of information gathered from such diverse range of sensors into a coherent mechanism that makes predictions about interactions and the coupling between modelling high level behaviour of people in a scene with signal processing issues of sensor fusion and tracking. Bayesian inference provides the core architectural framework to carry out the above research via a rigorous mathematical framework. The PROMETHEUS database discussed in the present work was created in support of this research and provides the necessary test bed for development of algorithms and testing their performance.

The main differences between the PROMETHEUS database and other related data, which were collected in related projects, such as HERMES[2], CogVis[3], URBANEYE[4], CARETAKER[5] , SERKET[6], etc, is that here the main focus falls on human behaviours recognition, interpretation and prediction. The PROMETHEUS database covers multiple applications, with interactions between individuals and groups of people, as well as interactions between humans and objects. To our best knowledge there is not other database of such size and coverage, nor exist other publically available resources that match the sensor set and the content profile of the PROMETHEUS database.

2. Objectives and Database Development

In brief, the present paper offers a comprehensive description of the design and the implementation of the audio part of the database. This database is in support of the development and the evaluation of the probabilistic structures that identify, track and recognize human actions, as well as the recognition of individual and group behaviours. Moreover, the database includes broad coverage of atypical events and abnormal behaviours, such as aggression and fights in urban environment, which are in support to developers of autonomous surveillance applications, as well as a variety of danger and crisis events in smart-home environment, which are useful for technology development in support of remote health-care applications. The broad-shouldered database design, the various recording scenarios, setups, and conditions, allow for embracing a range of real-world applications, including smart-home and human-robot interaction interfaces, security applications, such as: in-doors/outdoors public areas surveillance, airport terminals or city park supervision, etc.

[1] PROMETHEUS project (214901): http://www.prometheus-FP7.eu

[2] HERMES project (IST-2005-027110): http://www.hermes-project.eu/

[3] CogVis project (IST-2000-29375): http://cogvis.nada.kth.se/cogvis-home.html

[4] URBANEYE project (HPSE-CT2001-00094): http://www.urban-eye.net/

[5] CARETAKER project (IST FP6-027231): http://sceptre.king.ac.uk-/caretaker/

[6] SERKET project (ITEA): http://www.capvidia.com/files/SER-KET_PR.pdf

3. Database Design and Implementation

The database design is based on the requirements of the application scenarios mentioned in Section 2 and on the requirements evolving from the use of probabilistic framework for data processing and fusion. Specifically, the application scenarios defined the choice of test sites, which simulate the target environments, the contents of the action scripts, the number and contents of the task cards, the number of actors, the interactions between the actors, the interactions between actors and objects, etc. On the other hand, the technology requirements set the margins for the number of implementations of each script, the length of the individual recording sessions, the total size of the database, and least but not last the choice of sensor set for each setup. Other factors that were not controlled directly but which were accounted in the database design were the environmental conditions in the different setups (wind, noise, interferences from background activities, etc). Eventually we came up with a database design which had a more general nature than the strict requirements of the target applications and covers a wide range of indoors and outdoors scenarios.

The database consists of a number of recording sessions, which implement different aspects of the given application scenarios. All recordings belonging to a single scenario shared a common equipment setup, and represented a number of controlled conditions (for instance, in the smart-home application scenario one session was devoted to single-person actions, another for multiple-person interactions, etc). Each session was comprised of multiple action scenes concatenated in a single sequence, where each action scene was implemented a number of times, with different actors and different objects. The length of these sessions varied between 15 and 60 minutes. The recordings were performed with custom-build eight-channel microphone array and an eight-channel commercial microphone array.

Three sessions were recorded in order to fulfil the requirements of the airport scenario with average duration 18 minutes. Two sessions of 30 and 50 minutes duration respectively were found sufficient for the ATM - security scenario. Regarding the smart-home scenario three sessions of approximately 20 minutes average duration were captured. Lastly for the needs of the outdoors public security scenario, four sessions with total duration 76 minutes were implemented.

In total, the PPROMETHEUS database consists of more than 5 hours of recordings with both microphone arrays, but the portions where the audio is synchronized with the video cameras, 3D cameras and the thermal imaging cameras and other sensors are specifies as follows:

- Smart-home scenario – 62 min;
- Public security scenario (airport, outdoors) – 76 min;
- Public security scenario (airport, indoors) – 50 min;
- Public security scenario (ATM, outdoors) – 80 min.

The audio recorded by the custom build microphone arrays was sampled at rate of 32 kHz with 32-bits per audio sample and stored in WAV format. Furthermore, the single audio output of the commercial microphone array was sampled at 32 kHz with 16-bits per sample.

4. Annotation Procedure

The audio part of the database is annotated based on the circumstances in which the actors performed the respective actions as well as their emotional category. Additionally our main concern was to find efficient annotation tags in order to allow for the quick retrieval of the sound event. The annotation of speech and audio events is performed by using PRAAT[7], which provides a user-friendly environment while it is well suited for the needs since the audio sequence can be annotated in a number of different levels. Its output is an .xml file, which places particular focus on its simplicity, generality and usability. The corresponding video sequences were also given to the annotators to disambiguate certain situations and reduce the number of cases with disagreement among them. Throughout the creation of the corpus it was observed that when the video is present the annotators tend to reach consensus on nearly all controversial fragments. It should be mentioned that the corresponding image sequences are also manually tagged for serving person tracking and activity detection.

The following tiers and tags were used during the audio annotation phase:

(i) *Atypical Sound Event*, which referred to abnormal non-vocalic sounds with tags such as dropping of objects, fracture of material, footsteps, door sounds, fire alarm and other dominant events,

(ii) *Typical Sound Event*, which corresponded to normal/typical sounds, such as door bell, normal speech, interaction with Socrates (the smart-home assistant) etc,

(iii) *Background Noise*, which was edited when a background noise appears such as wind, speech in the background, music or other noise,

(iv) *Atypical Vocal Reactions*, which included human vocal sounds related to negative emotions while the tags were pain, fear, sorrow and anger,

(v) *Sex* (male/female), which was used only in one person typical and atypical speech,

(vi) *Verbal* (yes/no), used when speech audio events occur, and

(vii) *Audio Quality*, which reflected upon the quality of the audio signal with the next tags: clean, noise, music and other noise.

(viii) *Number of speakers*, which essentially is the number of active talkers at a given time.

The entire audio dataset was tagged by five annotators. Although the video sequences were provided, several audio events were not annotated with the same tags by all the members of the annotation team. In such cases we selected the tag that was used by the majority.

Figure 1 depicts a panic situation and a normal one captured during the recordings of the outdoors security scenario. In Figure 2 we show two representative

[7] Available at http://www.fon.hum.uva.nl/praat/

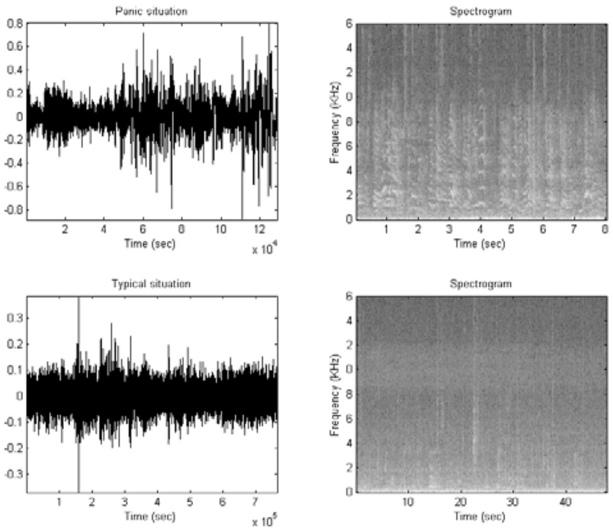

Figure 1: A panic situation (top row panels) and a typical one (bottom row panels) as regards to the outdoors security scenario.

examples including typical and atypical human sounds which were recorded during the recording sessions in the smart-home scenario.

5. Statistical Description of the Database

The final recordings contain rich information as regards different types of typical and atypical sound events, while a high degree of variation exist between samples

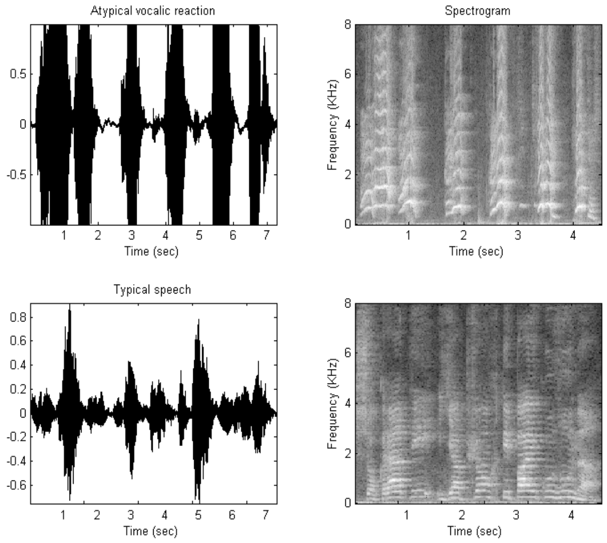

Figure 2: Characteristic waveforms and spectrograms derived from atypical (top row panels) and typical (bottom row panels) speech events, which were recorded during the smart-home scenario. The left-side panels present the speech waveforms and the right-side panels present the corresponding spectrograms.

of the same category. This came as a result of the database design which aimed at creating a multimodal dataset as close to real-life conditions as possible. A quantitative description of the sound events per scenario is tabulated in Table 1.

Scenario	Type of event	Acoustic event	Total number	Total duration (sec)
General purpose security scenario	abnormal	panic	10	128.6
		surprise	6	7.5
		anger	8	12.1
		scream	14	12.8
		people fighting	5	23.8
		pain	6	13.5
	normal	background speech	101	841.8
		normal speech	156	1233.2
		background noise (wind, birds, other noise)	21	255.1
Smart-home scenario	normal	normal speech	330	1790
		door bell	13	33
		background speech	25	693
		Socrates	59	224
	abnormal	fear	11	76
		surprise	9	10
		pain	6	19
		panic	12	81.3
		fire alarm	5	92
		fracture of material	6	12
		dropping of objects	16	16.2
ATM scenario	abnormal	anger	6	12.6
		panic	8	24.3
	normal	normal speech	43	81
		background speech	52	1607
		background noise (wind, birds, other noise)	16	185.2

Table 1: Acoustic events in the PROMETHEUS database

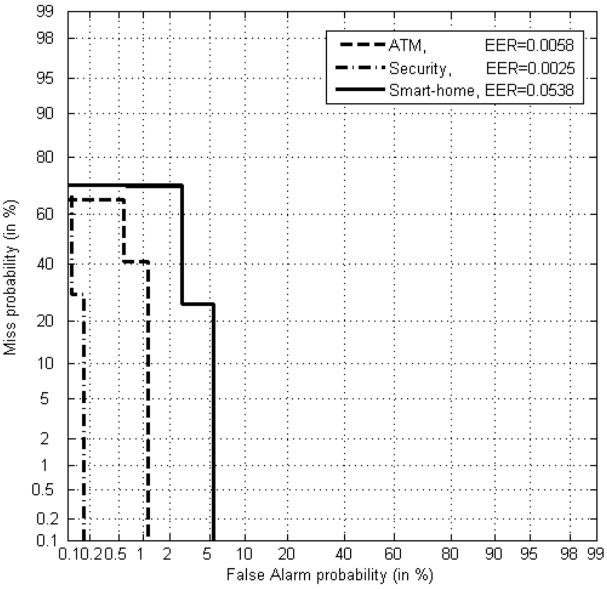

Figure 3: DET curves for all PROMETHEUS scenarios and the respective equal error rates. The target class is abnormal vocalic events.

In order to quantitatively evaluate the degree of difficulty that the PROMETHEUS database offers, we performed statistical and objective evaluation, exploring the discriminative capabilities of several acoustic parameters. We analyzed a number of potentially informative feature sets which can reveal the differences between typical and atypical sound events. Our primary target is to find a feature vector consisted of complementary coefficients in order to train statistical models towards building a robust classification system.

The following groups of acoustic parameters were employed: nonlinear Teager energy operator (TEO)-based features, critical band based TEO autocorrelation envelope area (TEO-CB-Auto-Env) features (Zhou et al., 2001), MPEG7 sound recognition descriptors (Casey, 2001), Mel-frequency cepstral coefficients, wavelet packet based audio descriptors (Sarikaya et al., 2000), pitch, duration, intensity, etc. A set of redundant parameters will also be evaluated since they may provide an efficient signal representation under the adverse conditions that were encountered during the PROMETHEUS recording sessions.

The above mentioned audio feature sets were tested for atypical sound event detection using Detection Error Trade-off (DET) curves (Martin et al., 1997). We utilized DET curves because traditional methods, such average recognition rate or confusion matrix do not present the twofold kind or error: fail to recognize an abnormal situation or detect one when it is not present. Both of them are crucial and should be taken under account for resulting with the best combination of features. The evaluation phase concerns the entire audio database including all the scenarios.

6. Experimentations on Detecting Atypical Situations

Lately the problem of unsupervised space monitoring based on the acoustic modality has gained quite a lot of attention from the signal processing community (e.g. Hiroaki et al., 2009). In order to deal with the variety of scenarios which were recorded in the PROMETHEUS database we have adapted the system explained in (Ntalampiras et al., 2009). This system was adapted towards detecting every type of atypical vocalic reaction that exists in the PROMETHEUS database (panic, scream, anger etc.) and the available sound events. In contrast to (Ntalampiras et al., 2009) here, we do not consider sound events such as explosion and gunshots, since they do not occur in the PROMETHEUS database. In brief, the sound recognition system has a hierarchical structure comprised of two stages: a) at the first stage the incoming sound is classified as vocalic or not vocalic and b) and in the second stage in the case a vocalic event is further processed to judge on its abnormality. .

Specifically, we trained diagonal GMMs for representing all the audio categories specified in Table 1. The following audio feature sets were used:

1. the first thirteen coefficients of the MFCC vector including the first one appended by their first derivatives,
2. TEO autocorrelation envelope area, pitch, pitch derivative and harmonicity to noise ratio.

The first group was used to differentiate between vocalic and non-vocalic sound events. The second one was combined with the first for classifying normal speech and atypical human expressions since it has the ability to capture the variations that intonation exhibits when speech signals are produced under abnormal circumstances. Details about the training of the system as well as results under different SNR conditions can be found in (Ntalampiras et al., 2009). In brief, fifty percent of the data was utilized for training the corresponding statistical models which the rest was employed for testing. The division train-test datasets was done in a random way. The performance of the first stage (vocalic/non-vocalic discrimination) was 100% for all scenarios. The performance of the second stage is presented as DET curves in Figure 3. Specifically, Figure 3 presents the DET curves for the smart home, security and ATM scenarios. The respective equal error rates are 5.38%, 0.25% and 0.58%. As we can see in the figure both the miss detection and false alarm probabilities have low values, which shows the good discriminative properties of the selected feature sets. We observe that the results on the outdoors recordings are better than the ones which belong to the indoors data. This might be due to the larger number of atypical events that exist in this scenario, including additional sound events such as *fracture of material*, *dropping of objects*, etc. We conclude that the described database is useful for creating probabilistic models which represent typical and hazardous situations under real-world conditions.

7. Discussion and Conclusion

This paper reported on the development of a new heterogeneous multi-sensor database that aims at supporting the research and development activities related to human behaviour tracking and interpretation. The database was recorded in two indoors (smart-home, public area) and two outdoors (security at ATM and airport) setups. In the present work we focused on the audio part of the database. The motivation behind the database design, as well as the implementation of the recording campaigns, the sensors used, the actual setups, and the data annotation tools and procedures were described. Sound recognition results are offered as reference for further research on the subject.

Besides support for the technology development within PROMETHEUS project, we deem that the PROMETHEUS database has the potential of becoming a widely-used resource, which could be used in some recently initiated related projects, such as INDECT[8], ADABTS[9], and others. We plan to make a major portion of the database publically available in the last quarter of year 2010.

8. Acknowledgements

This research was supported by the Prometheus project (FP7-ICT-214901), which is co-funded by the FP7 of the European Union.

9. References

Casey M. (2001). MPEG-7 sound-recognition tools. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), pp. 731--747.

Hiroaki N., Takanobu N. Hiroshi K. (2009). Acoustic-based security system: Towards a robust understanding of emergency shout. In 5th International Conference on Information Assurance and Security, Xian, China, 18-20 August.

Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M. (1997). The DET curve in assessment of detection task performance. *In 5th European Conference on Speech Communication and Technology*, Rhodos, Greece, 22-25 September.

Ntalampiras S., Potamitis I., Fakotakis N. (2009). An adaptive framework for acoustic monitoring of potential hazards. *EURASIP Journal on Audio, Speech and Music Processing*, Article ID 594103, doi:10.1155/2009/594103.

Sarikaya R., Hansen J.H.L. (2000). High resolution speech feature parameterization for monophone-based stressed speech recognition. *IEEE Signal Processing Letters*, 7(7), pp. 182--185.

Zhou G.-J., Hansen J.H.L., Kaiser J.F. (2001). Nonlinear Feature Based Classification of Speech Under Stress. *IEEE Transactions on Speech and Audio Processing*, 9(3), pp. 201--216.

[8] INDECT project (FP7-218086): http://www.indect-project.eu/

[9] ADABTS project (FP7-218197): http://ec.europa.eu/enterprise/-security/doc/fp7_project_flyers/adabts.pdf