THE SINS DATABASE FOR DETECTION OF DAILY ACTIVITIES IN A HOME ENVIRONMENT USING AN ACOUSTIC SENSOR NETWORK

Gert Dekkers[1,2,*], *Steven Lauwereins*[2,*], *Bart Thoen*[1,*], *Mulu Weldegebreal Adhana*[1,*], *Henk Brouckxon*[3,*], *Bertold Van den Bergh*[2,*], *Toon van Waterschoot*[1,2], *Bart Vanrumste*[1,2,4], *Marian Verhelst*[2], *Peter Karsmakers*[1]

[1] KU Leuven, Department of Electrical Engineering, Engineering Technology Cluster, Geel, Belgium.
[2] KU Leuven, Department of Electrical Engineering, Leuven, Belgium.
[3] Vrije Universiteit Brussel, Department ETRO-DSSP, Brussels, Belgium.
[4] IMEC, Leuven, Belgium.

ABSTRACT

There is a rising interest in monitoring and improving human well-being at home using different types of sensors including microphones. In the context of Ambient Assisted Living (AAL) persons are monitored, e.g. to support patients with a chronic illness and older persons, by tracking their activities being performed at home. When considering an acoustic sensing modality, a performed activity can be seen as an acoustic scene. Recently, acoustic detection and classification of scenes and events has gained interest in the scientific community and led to numerous public databases for a wide range of applications. However, no public databases exist which a) focus on daily activities in a home environment, b) contain activities being performed in a spontaneous manner, c) make use of an acoustic sensor network, and d) are recorded as a continuous stream. In this paper we introduce a database recorded in one living home, over a period of one week. The recording setup is an acoustic sensor network containing thirteen sensor nodes, with four low-cost microphones each, distributed over five rooms. Annotation is available on an activity level. In this paper we present the recording and annotation procedure, the database content and a discussion on a baseline detection benchmark. The baseline consists of Mel-Frequency Cepstral Coefficients, Support Vector Machine and a majority vote late-fusion scheme. The database is publicly released to provide a common ground for future research.

***Index Terms*—** Database, Acoustic Scene Classification, Acoustic Event Detection, Acoustic Sensor Networks

1. INTRODUCTION

There is a rising interest in smart environments to enhance the human experience and/or quality of life of its inhabitant. Such a system aims to understand the home scene to provide smart functionality, e.g. security, health monitoring [2] and entertainment using different types of sensors including microphones. In the context of Ambient Assisted Living (AAL) persons are monitored, e.g. to support patients with a chronic illness and older persons, by tracking their activities being performed at home [3, 4, 5, 6].

In order to make a smart home capable to automatically anticipate to forthcoming scenarios some form of sensing capabilities need to be available. Numerous sensor modalities have been investigated ranging from wearable [7] to contact-less sensors [8]. Existing research has been focussed either on a single modality or on the fusion of multiple modalities [6]. Compared to other modalities, microphone sensors are rarely used but contain highly informative data which can be exploited for multiple tasks [5]. Over the past decades, integrated components containing wireless radios and sensors are getting smaller in size, while maintaining computational power. This has led to using a network of sensors, which increases spatial sampling resolution. Therefore, this work focusses on using an Acoustic Sensor Network (ASN).

Another vital part of a smart home are the models that translate the data stream, acquired by the sensor(s), to information which can be used for a certain task. The task considered here is to detect an activity being performed, similar to the work in [3, 4, 5, 6]. When considering an acoustic sensing modality, an activity can be seen as an acoustic scene. The acoustic sensing literature has mainly covered the problems of Acoustic Event detection (AED) and Acoustic Scene Classification (ASC). An acoustic event is defined as a single consecutive event originated from a single sound source, e.g. a hand clap or a door knock. The ensemble of multiple events create a acoustic scenes describing a certain environment (e.g. a park or a living room) or, relevant to this paper, an activity being performed by a person (e.g. cooking or watching TV). Both the AED and ASC problems target the interpretation of the acoustic data. The rising interest in these problems has led to numerous public databases for a wide range of applications. The NAR dataset contains 41 sound events recorded by a humanoid robot Nao in a home environment [9]. The data used for the CLEAR 2006 and 2007 evaluations contain meeting room events collected by multiple microphones [10]. The DARES-G1 database contains annotations of sound events in different sound scenes, e.g. street and home [11]. The LITIS Rouen Audio Scene dataset [12], DCASE 2013 and 2016 databases [13, 14] consist of (binaural) recordings of events and/or scenes in public areas, e.g. office and park. The Multimodal subset of the SWEET-HOME database consists of recordings of daily activities performed by 21 different users leading to 26 hours of data. The recording setup consists of 7 microphone sensors, along with other sensor modalities, deployed in a smart home [6].

However, current databases do not possess all characteristics needed for our purposes: a) data collected in a home environment, b) activities being performed in a spontaneous manner, c) acquisition system based on an ASN, d) continuously recorded, and e) containing the activity when no person is present. Besides needing large databases to obtain accurate models and for algorithm validation, reference databases are important in algorithm development and comparison between algorithms.

[*] Thanks to VLAIO-SBO project Sound INterfacing through the Swarm (SINS) [1] for funding (contract 130006).

The main contributions of this paper are a) introducing a database, named *"SINS"*, of real-life recordings in a home environment using an ASN and b) providing a baseline detection benchmark as a reference to future work using this database. The paper is organized as follows: Section 2 introduces the recording environment and sensing hardware. Sections 3 presents the database content and recording procedure. This includes statistics about its content and how the annotation was performed. Section 4 describes the baseline detection benchmark and evaluation procedure. Section 5 shows the performance of the baseline along with an analysis. Finally, Section 6 presents conclusions and future work.

2. RECORDING ENVIRONMENT AND SETUP

The database was collected in a vacation home with a floor area of 50 m^2. The home consisted of five different rooms: a combined living room and kitchen, bathroom, toilet, bedroom and hall. Thirteen sensor nodes, each containing four microphones were distributed uniformly over the five rooms as indicated by Fig. 1. Details of the sensor's exact locations and height can be found in [15]. The sensor node has a modular design equipped with low-power audio sensing, audio processing and wireless capabilities [16]. The sensor node configuration used in this setup is a control board together with a linear microphone array. The control board contains an EFM32 ARM cortex M4 microcontroller from Silicon Labs (EFM32WG980) used for sampling the analog audio. The microphone array contains four Sonion N8AC03 MEMS low-power (±17μW) microphones with an inter-microphone distance of 5 cm. Although not used in this work, the setup can be used for sound source localization [16]. The sampling for each audio channel is done sequentially at a rate of 16 kHz with a bit depth of 12. The acquired data is sent to a Raspberry Pi 3 for data storage. The data is stored in chunks of one minute and timestamped. Timestamps were obtained based on an NTP protocol for rough synchronization, between the sensor nodes, with a sample accuracy of ∼500 ms. For algorithms demanding a more precise synchronization, an internal counter value of the control board is stored. The value was resetted every second by a GPS/Clock module. Using these counter values, a more precise synchronization (sample accuracy approximately ∼25 μs) could be obtained using interpolation techniques.

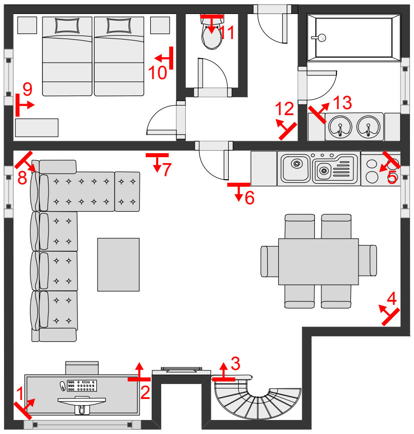


Figure 1: Floor map of the recording environment.

Room	Activity	Nr. ex.	duration (min.)
Living room	Phone call	22	8.17±13.73
	Cooking	19	16.62±9.49
	Dishwashing	15	6.37±1.49
	Eating	19	7.78±4.27
	Visit	9	13.3±12.11
	Watching TV	13	155.38±93.28
	Working	49	31.24±39.33
	Vacuum cleaning	13	4.79±2.14
	Other	200	0.75±0.95
	Absence	72	66.37±130.30
Bathroom	Drying with towel	10	1.67±0.28
	Shaving	13	1.91±1.46
	Showering	10	6.11±2.38
	Toothbrushing	19	1.41±0.25
	Vacuum cleaning	9	0.87±0.59
	Other	75	0.42±0.4
	Absence	35	248.56±263.62
Hall	Vacuum cleaning	9	3.31±1.11
	Other	164	0.36±0.22
	Absence	175	50.17±102.52
Toilet	Toilet visit	21	4.74±3.24
	Vacuum cleaning	7	0.53±0.07
	Absence	31	282.75±263.19
Bedroom	Dressing	28	1.53±1.10
	Sleeping	7	348.43±130.73
	Vacuum cleaning	7	1.04±0.27
	Other	22	0.27±0.23
	Absence	22	122.28±157.43

Table 1: Recorded activities for each room.

3. DATABASE CONTENT AND RECORDING PROCEDURE

One person lived in the environment for a continuous duration of one week. In order to have an as realistic as possible data recording there was no predefined set of scenarios that were simulated by some actor. Consequently, the recorded scenarios included being absent (e.g. getting groceries and going for a walk) or even receiving visitors. Although there was no restriction on the activities being performed, the number of activities that were labeled was limited as indicated in Table 1. In total 16 different activities were annotated in five different rooms. Table 1 lists the different activities along with the amount of examples and the mean and standard deviation of the duration of all examples for each room. Most of the activities are self-explanatory, except for "Working" and *"Other"*. "Working" contains recordings of the person doing work on a computer. The activity *"Other"* represents the presence of a person when not doing any activity of the ones listed in Table 1. Examples of recordings that are included in the *"Other"* activity are transitions between activities or the time between entering the room and starting an activity. In case of the *Living room* also sitting on the sofa or any other activity not listed in Table 1 was assigned to the *"Other"* category. In the case of the *Hall* this refers to crossing between rooms. Overall the database is strongly unbalanced, which indeed reflects the imbalance of different activities in daily life. In the case of the *Living room*, *"Absence"* and *"Watching TV"* are a factor 10 to 30 times larger in terms of total duration than the shortest activities *"Vacuum cleaning"* and *"Other"*. This ratio is even larger for the other rooms.

The annotation was performed in two phases. First, during the data collection a smartphone application was used to let the monitored person annotate the activities while being recorded. The person could only select activities listed in Table 1. The application was easy to use and did not significantly influence the transition between activities. Secondly, the start and stop timestamps of each activity were refined by using our own annotation software. In [15] more details are available on how these boundaries were chosen. During data collection we noticed occasional sensor node failure on two nodes. These were carefully annotated as well.

Postprocessing and sharing the database involves privacy-related aspects. Besides the person living there, multiple people visited the home. Also during the activity *"Phone call"*, one can partially hear the person on the other end. A written informed consent was obtained from all participants. The database and annotation are publicly available [15].

4. DETECTION BENCHMARK

The provided baseline is adopted from earlier work on a similar problem [4]. It consists of a Mel-Frequency Cepstral Coefficients (MFCC) feature extraction and a Support Vector Machine (SVM) based classifier. Each sensor node performs feature extraction and detection locally on the first out of four microphone signals. The obtained class label is fused centrally using majority vote to obtain a final class label. A decision is obtained for each room separately. In case of the *Living room* and *Bedroom*, decisions from respectively eight and two different sensor nodes are combined. In the other rooms no fusion is needed.

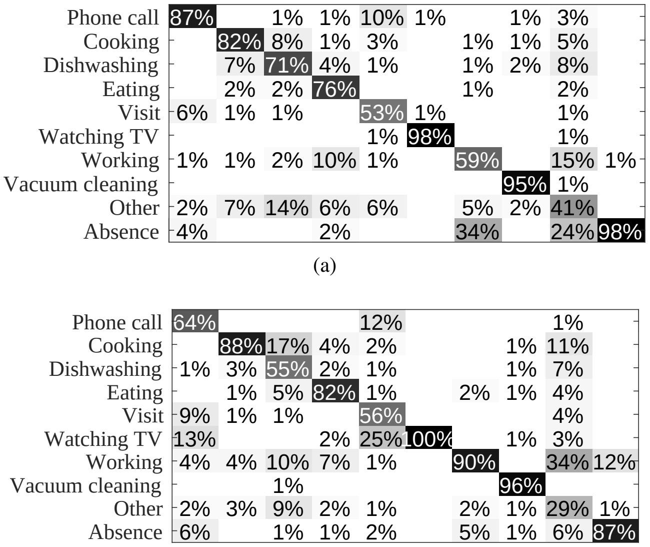
First, in each sensor node, the audio stream is transformed by a Short-Time Fourier Transform with a 30 ms hamming window and a 10 ms step size. Then, a mel-scale filterbank is used of length 26 with a frequency range of 500 to 8000 Hz. The mel-features are transformed to a lower dimension using Discrete Cosine Transform. The first 14 coefficients were kept, including the 0th order coefficient. Delta (Δ) and acceleration ($\Delta\Delta$) coefficients were also computed, based on a window length of 9 MFCC frames. Subsequently, the MFCCΔ/$\Delta\Delta$ feature vector stream was segmented using a sliding window of 15 s and a step size of 10 s. The window size is chosen based on the shortest average duration in Table 1. The mean and standard deviation are calculated for each feature dimension in the entire segment of 15 s which results in a total feature vector of length 84.

Finally, these features serve as an input to the model training and prediction phase of a SVM. SVM is a binary classifier that constructs a separating hyperplane such that the margin between two classes is maximized. For problems that are not linearly seperatable, a kernel maps the original space to a higher-dimensional space to make the separation easier. The kernel used here is the well-known radial basis function (RBF) kernel. To expand SVM to multi-class classification a 1-vs-1 coding scheme is used. The SVM hyper-parameters, kernel-bandwidth of the RBF and regularization parameter were tuned based on the training set [17]. Due to class imbalance, the contribution of each example in the model training phase is weighted based on class size.

The label assigned to the final feature vector is the active class at the middle of the segment window. Estimates of the current class therefore are based on non-causal information which introduces a delay of 7.5s in a practical setup. The feature vectors were grouped based on which example (Table 1) it belongs to. These groups were randomly assigned to a fold to be used for 4-fold cross-validation. F$_1$-score is used as a metric which does not take into account class imbalance. F$_1$-score's are obtained for each class separately and averaged to obtain the overall F$_1$-score.

5. RESULTS AND DISCUSSION

Results are analysed using normalized confusion matrices (nCM). For each room a confusion matrix is normalized by the marginals of either the column or the row. This provides insights into the precision and recall scores for each class and how the confusion is distributed. The precision is interpreted as how often the system is correct *when it estimates* a certain class, while recall provides insight into how often the system is correct *with respect to the ground truth.* In Fig. 2a classes are given on the y-axis. The precision (%) of each class is shown on the diagonal. The off-diagonals show the confusion with respect to a class on the diagonal in the same column. Therefore, the columns sum to 100%. For practical considerations, the values are rounded to the nearest integer. For example, in Fig. 2a, the class *"Working"* has a precision of 59% and is mostly confused with the class *"Absence"*. The analysis is the same for the recall, where confusion should also be looked up vertically in the same column.



Phone call	87%		1%	1%	10%	1%		1%	3%	
Cooking		82%	8%	1%	3%		1%	1%	5%	
Dishwashing		7%	71%	4%	1%		1%	2%	8%	
Eating		2%	2%	76%			1%		2%	
Visit	6%	1%	1%		53%	1%			1%	
Watching TV					1%	98%			1%	
Working	1%	1%	2%	10%	1%		59%		15%	1%
Vacuum cleaning								95%	1%	
Other	2%	7%	14%	6%	6%		5%	2%	41%	
Absence	4%			2%			34%		24%	98%

(a)

Phone call	64%				12%				1%	
Cooking		88%	17%	4%	2%			1%	11%	
Dishwashing	1%	3%	55%	2%	1%			1%	7%	
Eating		1%	5%	82%	1%		2%	1%	4%	
Visit	9%	1%	1%		56%				4%	
Watching TV	13%			2%	25%	100%		1%	3%	
Working	4%	4%	10%	7%	1%		90%		34%	12%
Vacuum cleaning			1%					96%		
Other	2%	3%	9%	2%	1%		2%	1%	29%	1%
Absence	6%		1%	1%	2%		5%	1%	6%	87%

(b)

Figure 2: Living room - (a) precision nCM - (b) recall nCM

Fig. 2 shows the output for the *Living room*. The averaged F$_1$-score over all classes is 82.3±2.2%. The worst performing class is *"Other"* which gets confused with *"Working"* and *"Absence"*. This seems logical because these three classes contain a great amount of silence. The class *"Absence"* contains audio when *"Vacuum cleaning"* is active in other rooms. The rooms in the vacation home were poorly acoustically isolated from each other and the door between *Hall* and *Living room* was partially opened due to the electricity cable of the vacuum cleaner. This, however, did not lead to high confusion between the two classes. The best performing classes are *"Absence"*, *"Watching TV"*, *"Vacuum cleaning"* and *"Cooking"* with F$_1$-scores around 95%.

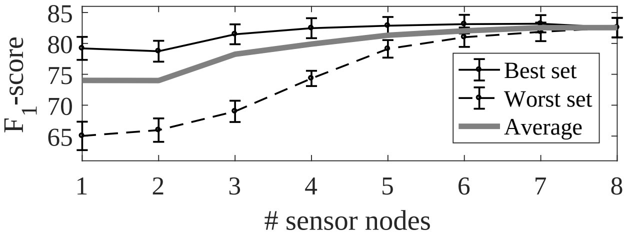
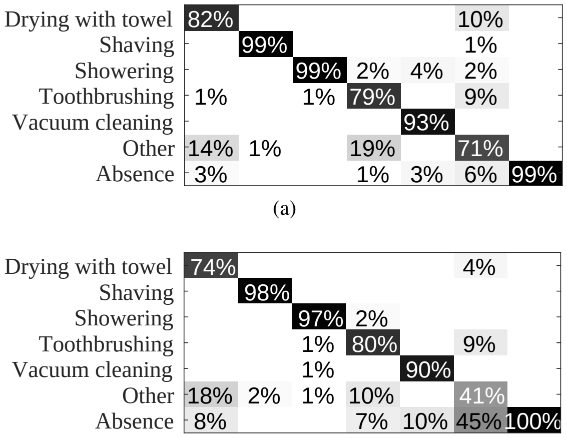


Figure 3: F_1-score versus amount of nodes used in the *Living room*

Fig. 3 shows the F_1-score with respect to the amount of nodes used. All possible sets of node combinations are tested. The best and worst set is shown together with the averaged performance over all sets. The gain in F_1-score, between using a single sensor or eight sensors, ranges between 3.4% and 17.3% depending on which sensor node is selected. On average the gain is 8.5%.

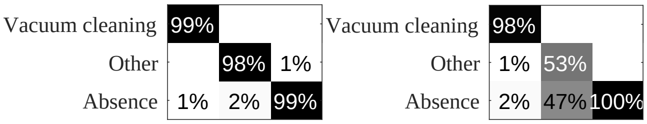


(a)

(b)

Figure 4: Bathroom - (a) precision nCM - (b) recall nCM

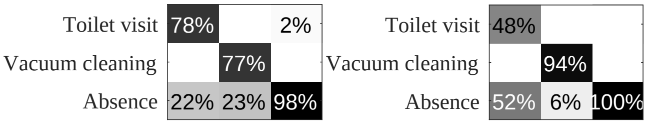
Fig. 4 shows the output for the *Bathroom*. averaged F_1-score over all classes is 84.8±2.0%. Similar trends are noticable here compared to the *Living room*. The classes *"Shaving"*, *"Showering"* and *"Absence"* perform above 95%. The worst performing classes are *"Drying with towel"* and again *"Other"*. Most of these classes are, as expected, confused with *"Absence"* and *"Other"*.



(a) (b)

Figure 5: Hall - (a) precision nCM - (b) recall nCM

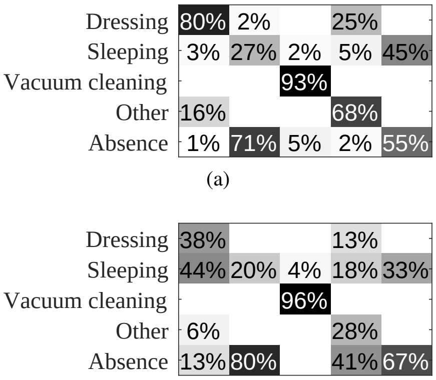
The output for *Hall* is shown in Fig. 5. The averaged F_1-score over all classes is 89.1±1.9%. The recall of class *"Other"* (53%) is considerably lower than the precision (98%). This shows that in case the system detects the class *"Other"* is active it is 98% correct, while when it should be *"Other"* is often confused with the class *"Absence"*. This could be due to the relatively short duration of the class *"Other"*. The average duration is 21.6 s (Table 1), while detections are based on segments of 15 s. In the case of *Hall*, the class *"Other"* only contains transitions between rooms.



(a) (b)

Figure 6: Toilet - (a) precision nCM - (b) recall nCM

Similar trends are observable for the results of *Toilet* in Fig. 6. The score of *"Vacuum cleaning"* is lower compared to other rooms due the shorter duration (31.8 s on average). The precision of *"Toilet visit"* (78%) is much higher than the recall (48%). A toilet visit is often without making any audible audio, causing the confusing with *"Absence"* for the recall nCM. When an event occurs however, it shows that these events often are recognized correctly leading to relatively high score. F_1-score over all classes is 79.0±8.6%.



(a)

(b)

Figure 7: Bedroom - precision and recall CM

The output for *Bedroom* is shown in Fig. 7. The overall F_1-score over all classes is 52.5±3.5%. This is the lowest overall F_1-score over all rooms, as expected due to the class *"Sleeping"*. The classes *"Dressing"*, *"Sleeping"*, *"Other"* and *"Absence"* are often confused between eachother. The only class performing above 90% is *"Vacuum cleaning"*.

6. CONCLUSIONS

In this paper we a) introduced the *"SINS"* database, a real-world database for detection of daily activities in a multi-room home environment using an acoustic sensor network and b) provided a first analysis on the detection performance using a benchmark system. The best performing room was the *Hall* with an F_1-score of 89.1%. The worst performing room is the *Bedroom* with an F_1-score of 52.5%. Both the database and annotation is available for download [15]. Future work will focus on a) improving the benchmark system and b) extending this database with isolated acoustic events and annotations on a sound event level using the acoustic scene database. For both subsets it is foreseen to also provide a benchmark and make the data public.

7. REFERENCES

[1] SINS. Sound INterfacing through the Swarm. [Online]. Available: http://www.esat.kuleuven.be/sins/

[2] F. Erden, S. Velipasalar, A. Z. Alkar, and A. E. Cetin, “Sensors in assisted living: A survey of signal and image processing methods,” *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 36–44, March 2016.

[3] L. Vuegen, B. Van Den Broeck, P. Karsmakers, H. Van hamme, and B. Vanrumste, “Automatic monitoring of activities of daily living based on real-life acoustic sensor data: A preliminary study,” in *Proc. Fourth workshop on speech and language processing for assistive technologies (SLPAT)*, 2013, pp. 113–118.

[4] ——, “Energy efficient monitoring of activities of daily living using wireless acoustic sensor networks in clean and noisy conditions,” in *2015 23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 449–453.

[5] M. Vacher, F. Portet, A. Fleury, and N. Noury, “Development of audio sensing technology for ambient assisted living: Applications and challenges,” *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 2, no. 1, pp. 35–54, January 2011.

[6] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, “The Sweet-Home speech and multimodal corpus for home automation interaction,” in *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014, pp. 4499–4506. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00953006

[7] O. D. Lara and M. A. Labrador, “A survey on human activity recognition using wearable sensors,” *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1192–1209, Third 2013.

[8] G. Laput, Y. Zhang, and C. Harrison, “Synthetic sensors: Towards general-purpose sensing,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’17. New York, NY, USA: ACM, 2017, pp. 3986–3999.

[9] M. Janvier, X. Alameda-Pineda, L. Girin, and R. Horaud, “Sound representation and classification benchmark for domestic robots,” in *Proc. IEEE International Conference on Robotics and Automation (ICRA14)*, May 2013.

[10] A. Temko, C. Nadeu, D. Macho, R. G. Malkin, C. Zieger, and M. Omologo, “Acoustic event detection and classification,” in *Computers in the Human Interaction Loop*, ser. Human-Computer Interaction Series. Springer, 2009, pp. 61–73.

[11] M. Grootel, T. Andringa, and J. Krijnders, “DARES-G1: Database of Annotated Real-world Everyday Sounds,” in *Proceedings of the NAG/DAGA Meeting 2009*, Rotterdam, 2009.

[12] A. Rakotomamonjy and G. Gasso, “Histogram of gradients of time-frequency representations for audio scene classification,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 1, pp. 142–153, Jan. 2015.

[13] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events,” *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[14] A. Mesaros, T. Heittola, and T. Virtanen, “TUT database for acoustic scene classification and sound event detection,” in *2016 24th European Signal Processing Conference (EUSIPCO)*, Aug 2016, pp. 1128–1132.

[15] KU Leuven, AdvISe research group. (2017) Datasets. [Online]. Available: http://iiw.kuleuven.be/onderzoek/advise/datasets

[16] B. Thoen, G. Ottoy, F. Rosas, S. Lauwereins, S. Rajendran, L. De Strycker, S. Pollin, and M. Verhelst, “Saving energy in WSNs for acoustic surveillance applications while maintaining QoS,” in *Proc. Sensors Applications Symposium (SAS)*, March 2017, pp. 1128–1132.

[17] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.