Semi-automated Annotation of Audible Home Activities

M. Garcia-Constantino*, J. Beltran-Marquez†§, D. Cruz-Sandoval‡, I.H. Lopez-Nava†‡, J. Favela‡,
A. Ennis*, C. Nugent*, J. Rafferty*, I. Cleland*, J. Synnott* and N. Hernandez-Cruz*
*School of Computing, Ulster University, Jordanstown, United Kingdom.
†CONACyT (Consejo Nacional de Ciencia y Tecnología), Mexico City, Mexico.
‡CICESE (Centro de Investigación Científica y de Educación Superior de Ensenada), Ensenada, Mexico.
§IPN (Instituto Politécnico Nacional), Tijuana, Mexico.
*Corresponding author: m.garcia-constantino@ulster.ac.uk

***Abstract*—Data annotation is the process of segmenting and labelling any type of data (images, audio or text). It is an important task for producing reliable datasets that can be used to train machine learning algorithms for the purpose of Activity Recognition. This paper presents the work in progress towards a semi-automated approach for collecting and annotating audio data from simple sounds that are typically produced at home when people perform daily activities, for example the sound of running water when a tap is open. We propose the use of an app called ISSA (Intelligent System for Sound Annotation) running on smart microphones to facilitate the semi-automated annotation of audible activities. When a sound is produced, the app tries to classify the activity and notifies the user, who can correct the classification and/or provide additional information such as the location of the sound. To illustrate the feasibility of the approach, an initial version of ISSA was implemented to train an audio classifier in a one-bedroom apartment.**

***Index Terms*—Data Annotation, Activity Recognition, Data Collection, Smart Microphones**

I. Introduction

Data annotation refers to the process of segmenting and labelling any type of data (images, audio or text) typically for the purpose of Activity Recognition. Manual annotation not only requires expertise in the field, but is also a time-consuming process, even using semi-automated annotation tools. On the other hand, automating the segmentation and labelling stages of annotation requires computational techniques to identify the segments in which the data of interest is contained [23], [24], and a knowledge database that contains all the possible data to be labelled. Annotated datasets are typically used to train machine learning algorithms to then automatically label unseen data. In the area of Activity Recognition, activities can be represented in different ways and at different levels of granularity (i.e. entire activity or individual steps that are part of a main activity).

In particular, audio annotation has applications ranging from professional media production to medical applications, such as heart rate monitoring [2]. The data annotation process will depend on the technology used to collect the data. In the case of audio data annotation, the use of video and audio has been typically used because the video images provide context in which the audio sounds were produced. The main downside of using video images as part of audio annotation is the possible privacy intrusion in cases where human subjects are shown on video. Thus, it is important to consider a less privacy-intrusive alternative using just audio data to annotate sounds.

The ubiquity of commercial smart voice assistants (a combination of smart speakers, microphones, and artificial intelligence), such as Amazon Echo, Google Home or Apple HomePod, and their constant improvement based on their consumers usage has led to their popularization among consumers of varied age ranges. In the context of audio data annotation, while it is possible to use smartphones for the same purpose and their use is more widespread than smart voice assistants, they are not dedicated and for this reason may be limited in terms of quality of audio captured and audio reproduced.

It is estimated that around 47.3 million adults in the United States have access to a type of commercial smart voice assistant[1], and that number is expected to increase. One of the main advantages of using smart voice assistants or smart microphones is that they are hands free, which allows the users to do other activities while interacting with the devices using voice commands. On the other hand, one of the main challenges in the adoption of smart microphones is the awkwardness that the user can experience when interacting with the devices for the first time or in a public setting.

The main contribution of this paper is an approach on the use of smart microphones to annotate audible home activities in a semi-automated manner and results from a feasibility initial evaluation. The approach is implemented as an Intelligent System for Sound Annotation (ISSA). Currently only one person has been considered to perform the home activities and only one array of microphones has been used. The number of audible home activities is limited but relevant in terms of: (i) their common occurrence in the daily life of a person, and (ii) their use in the area of Activity Recognition.

The work presented in this paper contributes to the ongoing research on smart environments carried out at the Smart Environments Research Group (SERG)[2] at Ulster University and at CICESE[3]. As one of the applications, it is intended to

[1]https://voicebot.ai/

[2]https://www.ulster.ac.uk/research/institutes/computer-science/groups/smart-environments/about

[3]https://www.cicese.edu.mx/

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

integrate the approach presented to the Smart Home in a Box (SHIB) approach being researched at Ulster University [9], which integrates different types of sensors to monitor elderly and people with physical and mental impairments.

The remainder of the paper is organized as follows: Section II presents the related work in the area of activity recognition using audio sources. Section III introduces the approach for data annotation of audible activities using smart microphones by describing an application scenario and a description of an Intelligent System for Sound Annotation (ISSA). Section IV describes the Audio Classifier Component of ISSA and its stages. Section V presents a discussion on the results obtained. Finally, Section VI presents the conclusions.

II. Related Work

Data annotation is an important process to generate reliable and sufficiently large enough annotated datasets that can be used for Activity Recognition. Machine learning algorithms use annotated data to produce classifiers that can then be used to classify unseen data. In [6], the ubiquitous use of smartphones is considered to design and evaluate a mobile based app that prompts the user to annotate their own activity data. Some of the activities considered in [6] were: walking, standing still, riding a bus and jogging. Results indicated that using prompts contributed in obtaining good activity labelling accuracy comparable to that of human observers.

An interesting work presented in [7] discusses the opportunities and challenges of collecting a large scale, diverse annotated dataset for Activity Recognition. In this work, accelerometer data was collected from 141 participants, who performed 3 of 18 activities from 6 scenarios of daily living. Examples of the identified challenges are the requirement of data quality adequate for classification and outliers found in data while examples of opportunities are the improvement of recognition results and generalization.

The opportunities and challenges in the data collection and annotation presented in [7] can be considered in the context of audio data. The annotation of a large and diverse audio dataset can be used to improve recognition of audible home activities. Regarding to the approaches dedicated to recognize audible home activities, some works have focused on recognizing human voice [4], [14], music [10], kitchen sounds [22], bathroom sounds [11], water flow [12], or even certain human sounds, such as snoring [8].

In general, the work dedicated to recognize audio has focused mainly in the selection of appropriate set of features to discriminate the activities to be recognized and in the proposal of algorithms and methods to improve the classification results. The aim of other works is to improve the segmentation process that directly affects data annotation.

The most common techniques for extracting features from audio signals can be divided into: (i) time-domain features, such as Zero-Crossing Rate (ZCR) [1], [5], [11], [22], and (ii) frequency-domain features, such as Short-Time Energy (STE) [21], total spectrum power [25] or Mel-Frequency Cepstral Coefficients (MFCC) [11], [15], [17], [20], [26].

On the other hand, the most common inference algorithms used to classify audible home activities are Support Vector Machines (SVMs) [4], [11], [14], [25], Hidden Markov Models (HMM) [16], [17], [28], Gaussian Mixture Models (GMM) [1], [5], [10], Hierarchical Hidden Markov Models (HHMM) [18], Dynamic Time Warping (DTW) [26] and Random Forests [15].

Regarding works focused on annotating audio signals, a real-time audio segmentation scheme is presented in [27], in which audio recordings were segmented and classified into basic audio types such as: (i) silence, (ii) sounds with music components, and (iii) sounds without music components. The techniques used were signal thresholds in energy, Zero-Crossing Rate, fundamental frequency, and spectral peak tracks. This scheme achieved an accuracy rate of more than 90% for audio classification.

In [8] the authors present a method which allows the automatic monitoring of snoring characteristics, such as intensity and frequency, from audio data captured via a freestanding microphone. This method allows the identification of periods of snoring, while rejecting silence, breathing, and other sounds. Data of six subjects was segmented and annotated manually. The proposed system correctly identified snores with 82-89% accuracy.

A voice detection technique was proposed in [5], and consists of two modules: (i) classification, and (ii) detection. In the classification module, audio segments were labeled for their use in training models considering 17 audio classes. The audio classes were categorized/labeled to build non-speech models from a sound scene database and from real data, and a speech model was built using a corpus of 1.8 million words. Then, in the detection module, acoustic event sounds were removed from the continuously listening environment. The performance of the speech/non-speech discrimination was 94.91% using GMM and 89.89% using SVM.

While the related work presented in this section involves the use of microphones and other devices to collect and annotate sounds, to our best knowledge smart microphones or smart voice assistants have not been used in this manner. Hence, the novelty of the work presented in this paper.

III. Approach for data annotation of audible activities

The following scenario describes how the approach proposed can be used for the annotation of audible activities:

A. Scenario

Arnold is an older adult who lives alone and that has a smart microphone at home. He has been using the device for several months and has become accustomed to interact with it through voice commands. One day, after Arnold prepares himself a juice in the blender, the device asks Arnold -*"What was that sound?"*- to which Arnold responds, -*"Oh, I prepared juice in the blender"*-. The system then records features associated to the sound annotated as "blender sound". On another day, Arnold prepares his juice again and the device identifies that

sound as likely to be a blender sound, however, the system is unsure because there is background noise that affects the recognition process. Thus, the system recognizes that the blender sound is coming from a different location than the usual. Due to this uncertainty, the system asks Arnold -*"Were you preparing juice?"*- to confirm the activity, to which Arnold responds -*"Yes"*-. As the system annotates new samples of Arnold preparing juice in the blender with different background and sounds mixtures, it will be improving its ability to recognize this sound under different circumstances. Moreover, the system will strategically annotate different sounds related to Arnold's daily life so that they can be used to recognize Arnold's activities. Furthermore, the annotated data could be used to generate models that could escalate the system to other users under other scenarios. The ability to recognize the location of the sound sources within the environment helps to improve classification results and to map the layout of users' scenarios. This location recognition can be useful to provide more insight into users' activities, such as when the user blends juice in a different location.

B. Intelligent System for Sound Annotation

An Intelligent System for Sound Annotation (ISSA) solution was implemented using a smart microphone to support the annotation of audible activities. ISSA continuously listens to sounds and gets feedback from the user to label and annotate audible activities in real-time.

ISSA's architecture (Figure 1) is deployed in a Raspberry PI 3 B+ board and uses additional features provided from the commercial smart microphone, called MATRIX Voice. The MATRIX Voice board has an array of 7 MEMS (microelectromechanical systems) microphones which allow gathering high-quality audios. The libraries used to configure the features (sampling frequency, minimum threshold for detection, and duration) of the recordings were: MATRIX HAL (Hardware Abstraction Layer) and ODAS (Open embeddeD Audition System). Furthermore, using these libraries, it is possible to get additional information about the sounds detected by the microphones such as 3D spatial position, energy, and audios recorded from each microphone. The audio classifier was implemented using HMM and the audio features MEL-MBSES described in [3]. To build a voice assistant, an open-source platform called Snips was used. Using the tools of Snips such as text-to-speech, speech-to-text, and modeling of capabilities, a voice assistant was built whose main task is to gather information from the user to label the audible activities detected in a specific environment. Data processing by ISSA must be performed locally to prevent privacy issues. This version of ISSA uses pre-trained models to classify and recognize sounds, however for a future version these classification models should be updated automatically using the data gathered to expand the range of audible activities recognized by ISSA.

Figure 2 shows an overview of the interaction between ISSA's components, users, and the environment. The description of how ISSA operates to annotate a specific audible activity is presented below.

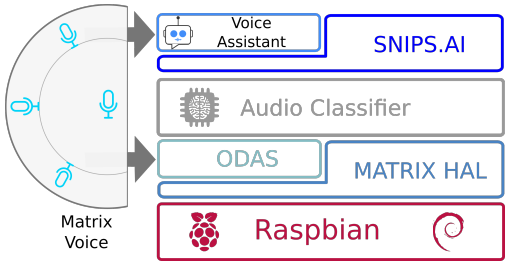

Fig. 1. Components and architecture of ISSA.

1) When ISSA detects sound energies above a specific threshold in a scale between 0 and 1 defined by the array of microphones (in this case a threshold of 0.75 was experimentally set), it begins recording the sound. The following criteria was defined to stop the recording: either the sound energy decreased below the threshold or the sound was recorded for a maximum of 10 seconds. The duration of 10 seconds was deemed as adequate because in general the sounds of the audible activities considered are short. In this stage, additional information on the detection -such as the 3D spatial position, timestamp, and duration- is obtained. ISSA ignores sounds whose energy do not overstep the defined threshold.
2) Once the sound is recorded, features are extracted and sent to the Audio Classifier Component of ISSA. The classifier tries to recognize the recorded sound as an audible activity in real-time based on a previous training.
3) A specific mode of the Voice Assistant will be triggered based on the result of the Audio Classifier. If it recognizes the sound recorded as an audible activity, then the Voice Assistant adopts the confirmation mode and asks the user to confirm its inference. If the Audio Classifier does not recognize the sound, then the Voice Assistant adopts the labeling mode and asks about the activity related to the sound recorded.
4) ISSA interacts with users via the Voice Assistant. Using Natural Language Processing (NLP), ISSA validates its inference or adds a new kind of audible activity using the confirmation or labeling mode respectively. In addition, ISSA asks about the semantic position of the activity; such as the bathroom, or kitchen. At any moment, the user can cancel the operation, and ISSA would delete all the information gathered for that instance.
5) Finally, ISSA stores the instance of the audible activity. The complete audio data instance (s_i) has the following structure:

$$s_i = \{\, activityName_i, audioChannel_i, \\ 3DPosition_i, semanticPosition_i, \\ timestamp_i\} \quad (1)$$

Where $activityName$ is the activity related to the sound, $audioChannel_i$ is an array with the 7 sounds

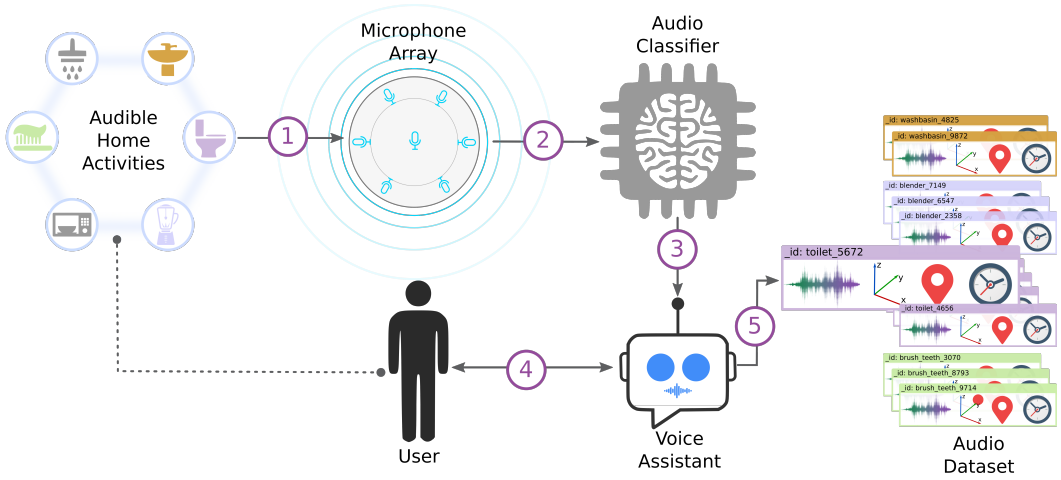

Fig. 2. Overview of the functionality of ISSA.

recorded by each microphone, $3DPosition_i$ is the spatial information according to the place where the sound was produced, $semanticPosition$ is the location provided by the user where the sound was detected, and $timestamp_i$ is the time when the sound was recorded.

The next section describes how the initial version of the Audio Classifier Component of ISSA was developed.

IV. Audio Classifier Component

A description of the process of creating the Audio Classifier Component of ISSA is presented, including: (i) data collection, (ii) model training, and (iii) classification results.

A. Data collection

The initial version of the Audio Classifier is limited to one-person/one-sound approach, *i.e.*, the assumption is that only one subject lives in this home. Audio produced by an individual was collected in a flat (see Figure 3), in which the microphones were placed at a central location with respect of the layout of the scenario. This location was considered as the place in which most audio produced at the apartment could be heard. A number of threshold configurations for the smart microphones to capture sounds were tested before collecting data to select the most optimal ones for this scenario.

Table I shows the sounds that were annotated and the locations in which they were produced. The selection criteria to consider these sounds was that they are typically associated to activities of daily living.

TABLE I
Audible home activities annotated and the locations in which they were produced. Living Room: LIV, Kitchen: KIT, Bathroom: BATH, Bedroom: BED.

Id	Sound	Location(s)
1	Running water from tap	KIT, BATH
2	Flushing toilet	BATH
3	Running water from shower faucet	BATH
4	Blender working	KIT
5	Unknown sounds	LIV, KIT, BATH, BED

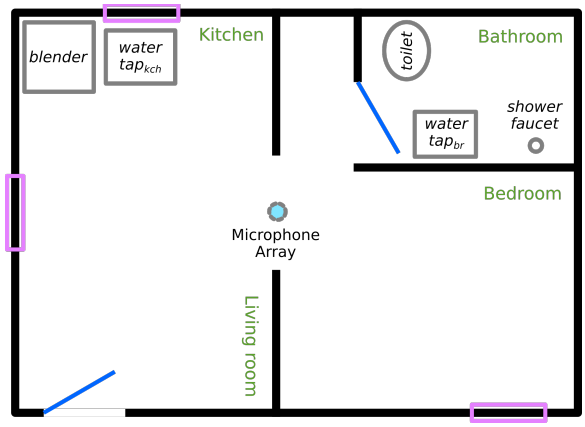

Fig. 3. Scenario used in this study.

Note that the audible home activity with Id5 is labelled as "Unknown sounds" and it refers to background sounds. The associated locations are all the ones considered in the scenario, see Figure 3. The annotation of sounds produced outside the apartment (street or neighbour flat/house) is not considered of interest for this paper. For the purposes of data annotation and data analysis, each of the audible home activity shown in Table I is associated to a class name.

B. Model training

Several samples for each of the audible home activities considered were collected and annotated (3 samples for the "running water from tap" sound, and 6 samples for the remaining sounds), resulting in a dataset of 27 samples, each with a maximum length of 5 seconds.

As defined in Equation 1, a data instance sample, for example for the "Flushing toilet" sound, is represented as:

$$s_i = \{\ flushingToilet, audioData, 3DPosition\ UNKNOWN, timestamp\} \quad (2)$$

In its current version, ISSA is not using the semantic location data, but eventually when the app detects that the

record has an "UNKNOWN" parameter, it will query the user about it. In this case, it asks the user about the location where the sound was produced. Overtime, it maps 3D coordinates to the related semantic location, so that at a later time, a similar record may have a similar form but it will include a question mark sign to the semantic position feature: "BATHROOM?". This will trigger the query: "Was that flushing toilet sound produced in the bathroom?". When the user confirms the location where the sound was produced, the record is changed to:

$$s_i = \{\ flushingToilet, audioData, 3DPosition \\ BATHROOM, timestamp\} \quad (3)$$

To train the model it was used the channel 8 delivered by the device, that is single signal after applying beamforming processing with the 7 channels. The audio collected and annotated was recorded at a sampling frequency of 44,100Hz, which is standard and of sufficient quality for data analysis. The analysis of the 27 samples of annotated data was carried out by extracting the audio features MEL-MBSES. Additionally, we trained HMMs (ergodic, 5 states) for each class.

C. Classification results

To evaluate, two audio streams with 224 seconds of continuous audio were collected, where each stream includes at least 1 sample of each class. The beamformed data was used for this evaluation. The objective of this data analysis was to verify that the audio dataset that resulted from the data annotation process is reliable for the automated data annotation of unseen realistic samples of audio. A process of continuous audio classification was carried out on audio excerpts of 5 seconds long with an overlap of 2 seconds.

Accuracy of 92.65% was obtained. Figure 4 shows how the excerpts from both streams were classified using the MEL-MBSES. As it can be seen, the classification is erroneous in the borders of the sounds and the background, these mistakes happen because in those places the time-window of 5 seconds includes audio of both classes. Note that the classification correctly identifies all audio events, however, improvements must be done to refine the segmentation. The confusion matrices show that water sounds are the sounds most confused with background sounds. Our results in this scenario aligns with current work which reports high accuracy results under the limitations stated in this experiment (sounds not overlapped), however, to attend challenges that arise in realistic scenarios we have to implement more sophisticated algorithms. Recent literature show that deep learning techniques are promising for sound event classification [13], [19]. In this sense, a large annotated dataset is paramount to improve classification, thus highlighting the necessity of annotation.

V. Discussion

The classification results obtained from the scenario using ISSA with one person performing one audible home activity at a time are promising. While this scenario may be deemed as a very simple case for which good results would be expected, it was required in order to have a solid foundation and baseline to be compared with more complex scenarios involving more people and sounds. The proposed modular architecture of ISSA has been designed to allow an efficient scaling of the system to address more complex scenarios.

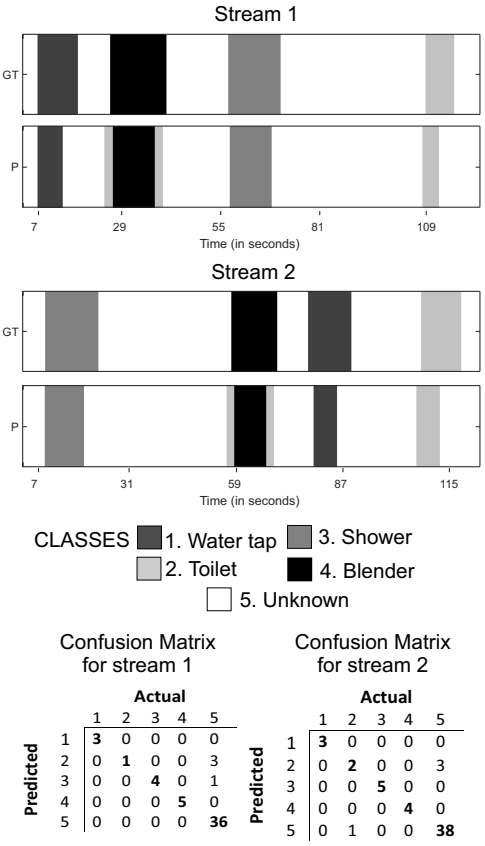

Fig. 4. Classification results. GT = Ground Truth, P = Predicted

The placement of the smart microphones in the environment scenario where the data annotations will be carried out will depend on the layout of the environment. As was mentioned in Section IV, in this case (see Figure 3) a number of configurations for the smart microphones to capture sounds were tested and the most optimal location was at the centre of the flat. Other locations in a scenario may not be suitable for sound annotation, such as a corner of a room or a table close to a wall, due to the limited range with respect to the total coverage area of the smart microphones. Locations next to walls or windows may introduce noise to sound annotations in the form of external sounds captured by the smart microphones.

VI. Conclusions

This paper has presented the work in progress towards a semi-automated approach for collecting and annotating audio data from simple sounds that are typically produced at home on a daily basis. In this work the data annotation was focused on the one-person/one-sound case.

The Audio Classifier developed had a low error rate, but it currently detects only a few types of sounds. The next step

will be to test the application in different settings to adjust the Voice Assistant so that it triggers at appropriate moments.

Future work will test ISSA using the current classifier with the feedback from the user as proposed in the scenario. Additionally, this work will explore the ability of the smart microphones to estimate the location of the sound source to assist in the annotation.

For future data collection and annotation, the cases of multiple users, more audible home activities and multiple smart microphones used in conjunction will be considered. Future work will also include transfer learning for training a classifier using the sounds from one scenario (flat/home) and apply it on another scenario to investigate how well it performs and how much training would it be required for an effective use. It is also planned to explore the case of using data fusion of the audio data annotations with other sensing sources of human activity, such as accelerometry. Finally, other possible applications outside of a home setting will be investigated, for example in Industry 4.0 for predictive maintenance to detect machinery malfunctions based on sounds.

ACKNOWLEDGMENT

Global Challenges Research Fund (GCRF) is acknowledged for supporting this project under the GCRF Building Capacity and Networking Grant.

REFERENCES

[1] P. K. Atrey, N. C. Maddage and M. S. Kankanhalli, "Audio based event detection for multimedia surveillance," In IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 5, 2006.

[2] M. Ballora, B. Pennycook, P. C. Ivanov, L. Glass and A. L. Goldberger, "Heart Rate Sonification: A New Approach to Medical Diagnosis," Leonardo, Vol. 37, No. 1, pp. 41–46, MIT Press, 2004.

[3] J. Beltrán, E. Chávez and J. Favela, "Scalable identification of mixed environmental sounds, recorded from heterogeneous sources," Pattern Recognition Letters, 68, pp. 153–160, 2015.

[4] S. H. Chen, S. H. Chen and B. R. Chang, "A Support Vector Machine Based Voice Activity Detection Algorithm for AMR-WB Speech Codec System," In 2nd International Conference on Innovative Computing, Information and Control, pp. 64–64, 2007.

[5] N. Cho and E. K. Kim, "Enhanced voice activity detection using acoustic event detection and classification," IEEE Transactions on Consumer Electronics, 57(1), 2011.

[6] I. Cleland, M. Han, C. Nugent, H. Lee, S. McClean, S. Zhang and S. Lee, "Evaluation of Prompted Annotation of Activity Data Recorded from a Smart Phone," Sensors, Vol. 14, No. 9, pp. 15861-15869, 2014.

[7] I. Cleland, M. P. Donnelly, C. D. Nugent, J. Hallberg, M. Espinilla and M. Garcia-Constantino, "Collection of a Diverse, Realistic and Annotated Dataset for Wearable Activity Recognition," IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 555-560, 2018.

[8] W. D. Duckitt, S. K. Tuomi and T. R. Niesler, "Automatic detection, segmentation and assessment of snoring from ambient acoustic data," Physiological measurement, 27(10), p. 1047, 2006.

[9] A. Ennis, J. Rafferty, J. Synnott, I. Cleland, C. Nugent, A. Selby, S. McIlroy, A. Berthelot and G. Masci, "A Smart Cabinet and Voice Assistant to Support Independence in Older Adults," International Conference on Ubiquitous Computing and Ambient Intelligence, pp. 466-472, 2017.

[10] S. Ghaemmaghami, "Audio segmentation and classification based on a selective analysis scheme," In 10th International Multimedia Modelling Conference, pp. 42–48), 2004.

[11] T. Giannakopoulos and G. Siantikos, "A ROS framework for audio-based activity recognition," In 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments, p. 41, 2016.

[12] P. Guyot, J. Pinquier, and R. Andre-Obrecht, "Water flow detection from a wearable device with a new feature, the spectral cover," In International Conference on Content - Based Multimedia Indexing, pp. 1–6, 2012.

[13] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss and K. Wilson, "CNN Architectures For Large-Scale Audio Classification," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131-135, 2017.

[14] Z. Huang, Y. C. Cheng, K. Li, V. Hautamaki and C. H. Lee, "A Blind Segmentation Approach to Acoustic Event Detection Based on I Vector," in 14th Annual Conference of the International Speech Communication Association, 2013.

[15] A. Kumar, R. Singh and B. Raj, "Detecting sound objects in audio recordings," In 22nd European Signal Processing Conference, pp. 905–909, 2014.

[16] L. Ma, B. Milner and D. Smith, "Acoustic environment classification," ACM Transactions on Speech and Language Processing, 3(2), pp. 1-22, 2006.

[17] A. Mesaros, T. Heittola, A. Eronen and T. Virtanen, "Acoustic event detection in real life recordings," In 18th European Signal Processing Conference, pp. 1267–1271, 2010.

[18] Y. T. Peng, C. Y. Lin, M. T. Sun and K. C. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," In IEEE International Conference on Multimedia and Expo, pp. 1218–1221, 2009.

[19] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Processing Letters, Vol. 24, No. 3, pp. 279-283, 2017.

[20] J. A. Stork, L. Spinello, J. Silva and K. O. Arras, "Audio-based human activity recognition using non-markovian ensemble voting," In 21st IEEE International Symposium on Robot and Human Interactive Communication, 2012.

[21] A. Tripathi, D. Baruah and R. D. Baruah, "Acoustic event classification using ensemble of one-class classifiers for monitoring application," In IEEE Symposium Series on Computational Intelligence, pp. 1681–1686, 2015.

[22] A. Vafeiadis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen and R. Hamzaoui, "Audio-based Event Recognition System for Smart Homes," In 14th IEEE International Conference on Ubiquitous Intelligence and Computing, 2017.

[23] H. Vereecken, J.-P. Martens, C. Grover, J. Fackrell and B. Van Coile, "Automatic Prosodic Labeling of 6 Languages," In 5th International Conference on Spoken Language Processing, Vol. 4, pp.1399-1402.

[24] A. Vorstermans, J.-P. Martens and B. Van Coile, "Automatic Segmentation and Labelling of Multi-Lingual Speech Data," Speech Communication, Vol. 19, No. 4, pp. 271-293, Elsevier, 1996.

[25] K. Yatani and K. N. Truong, "BodyScope: a wearable acoustic sensor for activity recognition," In ACM Conference on Ubiquitous Computing, pp. 341–350, 2012.

[26] Y. Zhan, J. Nishimura and T. Kuroda, "Human activity recognition from environmental background sounds for wireless sensor networks," IEEJ Transactions on Electronics, Information and Systems, 130(4), pp. 565-572, 2010.

[27] T. Zhang and C. C. J. Kuo, "Heuristic approach for generic audio data segmentation and annotation," In 7th ACM international conference on Multimedia (Part 1), pp. 67–76, 1999.

[28] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson and T. S. Huang, "Real-world acoustic event detection," Pattern Recognition Letters, 31(12), pp. 1543–1551, 2010.