homeSound: A High Performance Platform for Massive Data Acquisition and Processing in Ambient Assisted Living Environments

Marcos Hervás[1], Rosa Ma Alsina-Pagès[1] and Joan Navarro[2]

[1]*GTM - Grup de Recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull, Barcelona, Spain*

[2]*GRITS - Grup de Recerca en Internet Technologies & Storage, La Salle - Universitat Ramon Llull, Barcelona, Spain*

Keywords: Ambient Assisted Living, Sensor Network, Machine Hearing, Audio Feature Extraction, Machine Learning, Graphics Processor Unit.

Abstract: Human life expectancy has steadily grown over the last century, which has driven governments and institutions to increase the efforts on caring about the eldest segment of the population. The first answer to that increasing need was the building of hospitals and retirement homes, but these facilities have been rapidly overfilled and their associated maintenance costs are becoming far prohibitive. Therefore, modern trends attempt to take advantage of latest advances in technology and communications to remotely monitor those people with special needs at their own home, increasing their life quality and with much less impact on their social lives. Nonetheless, this approach still requires a considerable amount of qualified medical personnel to track every patient at any time. The purpose of this paper is to present an acoustic event detection platform for assisted living that tracks patients status by automatically identifying and analyzing the acoustic events happening in a house. Specifically, we have taken benefit of the amazing capabilities of a Jetson TK1, with its NVIDIA Graphical Processing Unit, to collect the data in the house and process it to identify a closed number of events, which could led doctors or care assistants in real-time by tracking the patient at home. This is a proof of concept conducted with data of only one acoustic sensor, but in the future we have planned to extract information of the sensor network placed in several places in the house.

1 INTRODUCTION

Human life expectancy is increasing in the modern society (Suzman and Beard, 2015). This drives our society to face new challenges in terms of health care because the number of patients to attend is increasing according to (National Institute on Aging, 2007; Chatterji et al., 2008) the people ageing who need support (Lafortune and Balestat, 2007). Nowadays, public and private health services try to avoid long term hospitalizations and, instead, foster the elderly to remain at home for two reasons: on the one hand, it is better for their health to keep them—while not suffering from severe deterioration—in their own environment and, on the other hand, it is much cheaper for health services. However, there is still a quality gap between the service provided at medical facilities and the service provided at patients' home.

Technology is a powerful tool that can contribute to address this problem by enabling medical staff to monitor and attend patients while they are at home. Ambient Assisted Living (AAL) (Vacher et al., 2010) reduces the personnel costs in health assistance. AAL consists of monitoring the preferred living environment of the patients with intelligent devices that can track their status and improve their life quality. To address this hot research topic, several engineering projects have been proposed to discuss the feasibility of deploying smart robots at the home of elderly not only to cover routine tasks, but also to remind them to have their medication or interact with them through serious games (Morsi and Shukla, 2015). One of the main challenges that this proposals open is the huge amount of data that these robots have to collect in order to provide a meaningful response for patients. Typically, these robots have limited computing capabilities and, thus, are able to process data from a reduced number of sensors.

This paper explains the proof of concept of a software and hardware platform designed to recognize a set of the predefined events from the environmental sound in a house. This information can be later used to infer the in-home context and detect some situations of risk. To process data from several sources

HervÃąs M., Alsina-PagÃĺs R. and Navarro J.
homeSound: A High Performance Platform for Massive Data Acquisition and Processing in Ambient Assisted Living Environments.
DOI: 10.5220/0006209701820187
In *Proceedings of the 6th International Conference on Sensor Networks (SENSORNETS 2017)*, pages 182-187
ISBN: 421065/17

(e.g., microphones) and conduct the computations associated to audio event identification in parallel, the system implements a recognition scheme using a NVIDIA Jetson TK1 (NVIDIA Corp., 2016) Graphical Processing Unit (GPU). This platform can reach to several decisions depending on the situation and home, and the final conclusion can be activating some kind of alarm or just track the patients behaviour for health purposes. Overall, the purpose of this work is to present the first approach to the implementation of an acoustic event recognition platform and the obtained results when classifying a limited corpus of events.

The reminder of this paper is organized as follows. Section 2 reviews the related work on environmental sound recognition; it is specially focused on ambient assisted living environments. Section 4 elaborates on the technical details of the proposed algorithm to solve the problem, which corresponds to a basic implementation. Section 3 gives details about the selected platform and its convenient features to process audio data. Section 5 describes the algorithm used to classify the events and shows the obtained results when running on the chosen platform. Finally, Section 6 detail the conclusions and future work of this project.

2 RELATED WORK

There are several approaches in the literature that aim to extract features from the sound. From these features, it is possible to create a corpus of a close universe of different sounds and train a machine learning system to classify the source of the sound. Therefore, environmental sound recognition has emerged as a hot research topic today, which has led to some interesting applications (Chachada and Kuo, 2014); from animal recognition to surveillance, including ambient assisted living use cases.

Interest in detecting in-home sounds started from the beginning of this technology in 2005. Chen et al. (Chen et al., 2005) was monitoring the bathroom activity using only the sound information. Afterwards, with research not detailed in this work, robust environment sound recognition motors were designed in 2008 (Wang et al., 2008). One of the most challenging problems to be solved in this field, is to take into account the varying acoustic background, the noise sources. In this regard, the project SonicSentinel (Hollosi et al., 2011) uses noise-robust model-based algorithms to evaluate the noise sources. Evolving this technology, Valero et al. (Valero and Alías, 2012) succeeded on classifying audio scenes. Additionally, several works can be found about audio analysis in a smart home to help doctors on the early diagnose of dementia diseases for the elder (Guyot et al., 2013). Also, it is worth mentioning that conditional random fields have been used to build an event detection framework in a real-world environment of eight households (Matern et al., 2013), which led the system to be sometimes unreliable.

From the applications point of view, one of the most popular use-cases nowadays of audio event recognition is its use in the smart home (Chan et al., 2008), especially when conceiving systems to meet the needs of the elderly people. The constraints around the design of a smart home for health care (Vacher et al., 2010) based on audio event classification are as follows: *i)* the degree of dependency of the disabled person, *ii)* the quality of life to be improved by means of automatizing the processes, and *iii)* the distress situations recognition and the activation of the preassigned protocols, including reducing the false alarm situations (Goetze et al., 2012).

Despite the fact that there are several solutions in the literature (Vacher et al., 2013) that consider these three constraints, the primary goal of the platform presented in this paper is to accurately address the third one. Additionally, our proposal aims to meet the needs of ambient assisted living, which are the following (van Hengel and Anemüller, 2009): *i)* increasing the comfort of living at home, *ii)* increasing the safety, through detecting dangerous events and *iii)* supporting health care by professionals, through detecting emergencies and monitoring vital signs.

3 SYSTEM DESCRIPTION

The proposed system diagram to monitor audio events in ambient assisted living environments is shown in Figure 1.

As far as the proof of concept herein presented is concerned, the system relies on a network of microphones consistently deployed around the house (see Figure 2). The microphones are installed in such a way that they provide the maximum entropy of a given event (i.e., it is not necessary to analyze together different audio sources).

The microphones used in this application to sense the environmental sound should present a good trade-off between the frequency response and cost, for this reason tests are being conducted with the electret condenser microphone CMA-4544PF-W (CUI inc., 2003) of the manufacturer CUI inc. with a very low price.

In this way, each microphone transmits sounds to

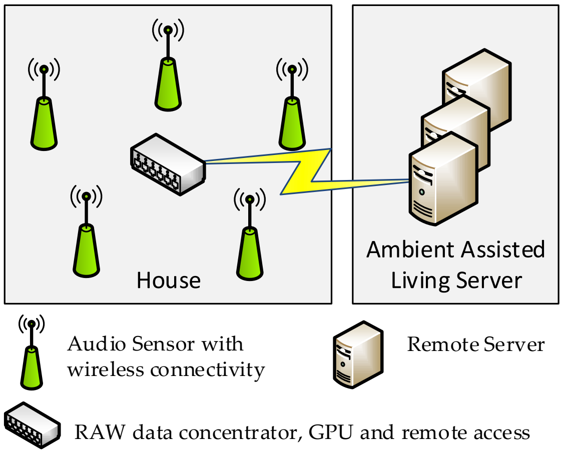

House
Ambient Assisted Living Server
Audio Sensor with wireless connectivity
Remote Server
RAW data concentrator, GPU and remote access

Figure 1: Block diagram of the network elements of this system.

this device that acts as a concentrator—the core element of our proposal. As a matter of fact, this concentrator *i)* collects all the audio sounds of the house, *ii)* processes them in order to extract their features, *iii)* infers the source of the audio event, and *iv)* sends this information to a remote server that monitors the needs of the people living in the house.

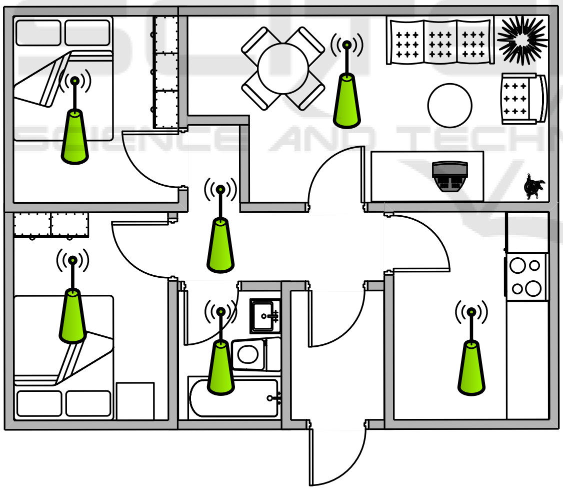


Figure 2: Example of the proposed audio sensors network deployed in a house.

The concentrator platform used in this work is the NVIDIA Jetson TK1 developer kit. This platform is based on the Tegra K1 SOC, which is composed of *i)* an NVIDIA Kepler GPU with 192 CUDA cores and *ii)* a quad core ARM cortex-A15 CPU. The Tegra family is the proposal of the NVIDIA manufacturer for mobile processors in which you need GPU-accelerated performance with low power consumption.

This GPU is able to process up to 192 threads in parallel. Kepler architecture offers an improvement of performance up to 3 times more than the previous version, Fermi, (NVIDIA Corp., 2014). This level of concurrency allows us to process audio events of several sources in real-time.

Therefore, to exploit the parallel capabilities of the concentrator, it opens a thread to process each audio source and infer the event that generated every sound.

4 MACHINE HEARING

Endowing machines with the ability of hearing the acoustic environment to detect and recognize an event as humans do, is known as machine hearing. The algorithm used in this work is based on *i)* feature extraction using mel-frequency cepstral coefficients (MFCC) (Melmstein, 1976) and *ii)* pattern recognition using the k-Nearest Neighbors classifier (KNN) (Cover and Hart, 1967), see Figure 3.

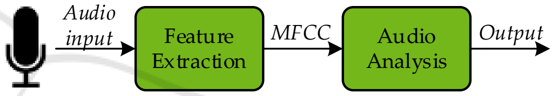


Figure 3: Block diagram of a Hearing Machine algorithm.

4.1 Feature Extraction

Feature extraction aims to obtain a representation of audio events in which the dimensionality of this parametrization is much lower than the original samples (Alías et al., 2016). This parametrization will be the input data of the classifier. The parametrization used in this work, MFCC (Melmstein, 1976), uses an approach based on perceptual-based frequency using the Mel scale (Liang and Fan, 2014).

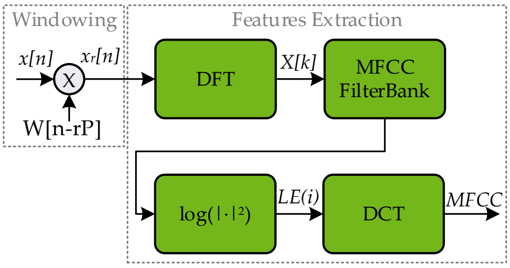


Figure 4: Block diagram of the feature extraction based on the Mel coefficients used in this work.

The incoming audio stream is divided into blocks of 30 ms with an sliding window. These frames are transformed into frequency domain using the DFT to measure the power of different bands of the spectrum. The power measures are conducted with a bank of 48 filters using the Mel scale (see Figure 5). The MFCC coefficients are obtained from the Discrete Cosine Transform (DCT) of the logarithm of these 48 values. The higher order coefficients of the DCT are discarded to obtain a reduced dimensionality characterization of the sound event, this compression can be done because the main information is in the low frequency components of the signal's spectral envelop. The final number of MFCC coefficients is 13.

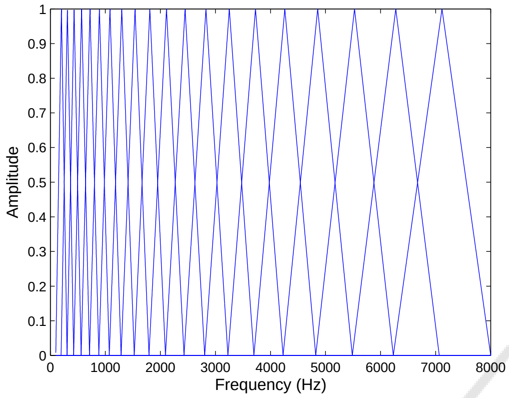


Figure 5: Example of a Mel scale with a filter-bank of 20.

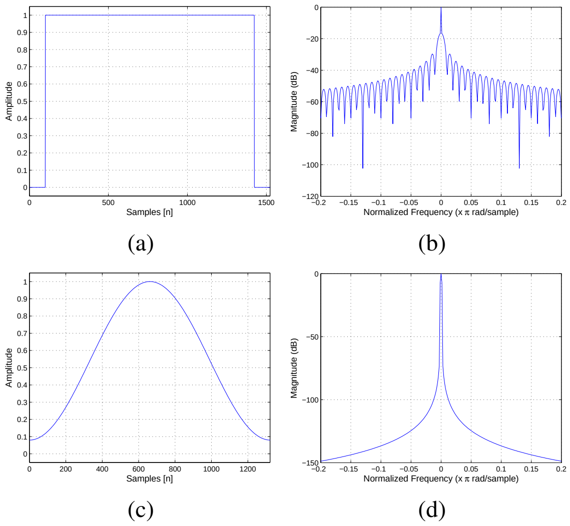
Window lengths between 10 and 50 ms are usually used to detect transient audio events (Fu et al., 2011). A Hamming windowing is also applied to this frame of samples to improve the frequency resolution in the Discrete Fourier Transform (DFT)—as we can see comparing the differences between square and Hamming windows in Figure 6. This sliding block has an overlap of 50% of samples to compensate the power reduction of the data blocks due to the laterals of the Hamming window, see Figure 6.

The Mel scale is a perceptual scale which aims to emulate the behaviour of the human hearing. As we can observe in Figure 5, Mel scale is a bank of triangular filters.

4.2 Automatic Audio Classification

Machine learning algorithms are widely used in the literature of speech technologies to automatically classify audio samples. In fact, most of the audio recognition systems settle the use of the MFCC coefficients as baseline in terms of feature extraction (Alías et al., 2016). Then, when the signal is processed and the features are already extracted, a k-Nearest Neighbors (kNN) (Cover and Hart, 1967) system can be run (Zhang and Zhou, 2005).

Hence, we have followed this approach and trained a kNN classifier as follows. We have built a training data set composed by 2850 audio samples belonging to 14 in-home events lasting a total number of 20 hours. We have split every sample in several sub samples as detailed above, and for every sub sample we have computed the MFCC coefficients. This results in a vector of 13 components (each one corresponding to its associated MFCC) for every sound sub sample. As a result, a sound sample is characterized with a set of vectors. The number of vectors that characterize a given sound depends on the length of the training sound, which would limit the classifier accuracy (i.e., shorter sounds of the same sound type would probably be misclassified. Therefore, to address this issue, we have built a bag of words with all the vectors belonging to the same sample using the aforementioned kNN. The resulting vector has a fixed length of K components. This gives an idea of how many portions of the training sound set belong to each centroid of the kNN, which at the same time removes the temporal dimension of the sound event. Next, we normalize all these resulting vectors to made the suitable for a fair comparison. With this fixed size set of normalized vectors, we finally train a Support Vector Machine (running on the concentrator platform).



(a)

(b)

(c)

(d)

Figure 6: Comparison between a squared and Hamming windows in time and frequency domain: *a)* is a squared window, *b)* is the spectrum of the squared window, *c)* is a Hamming window and *d)* is the spectrum of the Hamming window.

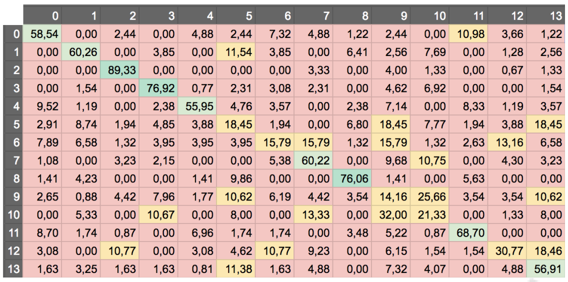
Finally, when our system is in exploitation mode, the concentrator platform builds extracts the audio sub samples and builds the fixed size vector accordingly. Then, this vector is delivered to the Support Vector Machine (SVM) to predict the event.

5 RESULTS

With the data set described in the previous section we have conducted our experimentation. Specifically, we have used 60% of instances to train the SVM and the other 40% to test it. To obtain statistically significant results we have performed a 10-fold cross validation.

We have found that this system is able to recognize the following events with an overall accuracy of 73%: someone falling down, slice, screaming, rain, printer, people talking, frying food, filling water, door knocking, dog bark, car horn, glass breaking, baby crying, water boiling.

Figure 7 shows the confusion matrix.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	58,54	0,00	2,44	0,00	4,88	2,44	7,32	4,88	1,22	2,44	0,00	10,98	3,66	1,22
1	0,00	60,26	0,00	3,85	0,00	11,54	3,85	0,00	6,41	2,56	7,69	0,00	1,28	2,56
2	0,00	0,00	89,33	0,00	0,00	0,00	0,00	3,33	0,00	4,00	1,33	0,00	0,67	1,33
3	0,00	1,54	0,00	76,92	0,77	2,31	3,08	2,31	0,00	4,62	6,92	0,00	0,00	1,54
4	9,52	1,19	0,00	2,38	55,95	4,76	3,57	0,00	2,38	7,14	0,00	8,33	1,19	3,57
5	2,91	8,74	1,94	4,85	3,88	18,45	1,94	0,00	6,80	18,45	7,77	1,94	3,88	18,45
6	7,89	6,58	1,32	3,95	3,95	3,95	15,79	15,79	1,32	15,79	1,32	2,63	13,16	6,58
7	1,08	0,00	3,23	2,15	0,00	0,00	5,38	60,22	0,00	9,68	10,75	0,00	4,30	3,23
8	1,41	4,23	0,00	0,00	1,41	9,86	0,00	0,00	76,06	1,41	0,00	5,63	0,00	0,00
9	2,65	0,88	4,42	7,96	1,77	10,62	6,19	4,42	3,54	14,16	25,66	3,54	3,54	10,62
10	0,00	5,33	0,00	10,67	0,00	8,00	0,00	13,33	0,00	32,00	21,33	0,00	1,33	8,00
11	8,70	1,74	0,87	0,00	6,96	1,74	1,74	0,00	3,48	5,22	0,87	68,70	0,00	0,00
12	3,08	0,00	10,77	0,00	3,08	4,62	10,77	9,23	0,00	6,15	1,54	1,54	30,77	18,46
13	1,63	3,25	1,63	1,63	0,81	11,38	1,63	4,88	0,00	7,32	4,07	0,00	4,88	56,91

Figure 7: Confusion Matrix. Events are ordered from left to right as follows: falling down, slice, screaming, rain, printer, people talking, frying food, filling water, door knocking, dog bark, car horn, glass breaking, baby crying, water boiling.

In this confusion matrix we can see how often the SVM misclassifies a given class and, thus, assigns a wrong event to an audio sample. It is shown that in general, the best results for each sample are obtained when testing the sound event against itself. Also, it depicts the skill of the classifier on distinguishing one audio event from the others. The optimal value of this confusion matrix should be an Identity Matrix with the value 100 on its diagonal.

Although the classifier performs reasonably well, it gets confused on some sound events that have several MFCC coefficients pretty similar. For instance, on row 6 in Fgure 7, door knocking, people talking and frying food have similar MFCC vector patterns and, thus, the SVM features a low accuracy in these specific situations. To address this concern, we plan to (1) complement the training vector set with other sources in addition to MFCCs, and (2) use a more sophisticated classifier such as a deep net.

6 CONCLUSIONS

Preliminary results of our paper encourages us to keep on working on the analysis of the events happening in the house. We will work with the feature extraction improvement with other methods, as well as we will test more machine learning algorithms to increase the accuracy of the system with just one acoustic measurement. Next steps after this proof of concept using the Jetson TK1 are the expansion of the platform, by means of using a wider sensor network, where several autonomous acoustic sensors sending data to the GPU to be processed. In this stage, an important part of the work will be focused on the optimization of the acoustic event detection algorithm to take advantage of the parallelization of the GPU unit.

ACKNOWLEDGEMENTS

The authors would like to thank the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement (Generalitat de Catalunya) under grant ref. 2014-SGR-0590.

REFERENCES

Alías, F., Socoró, J. C., and Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, 6(5):143.

Chachada, S. and Kuo, J. (2014). *Environmental sound recognition: A survey*. APSIPA Transactions on Signal and Information Processing.

Chan, M., Estève, D., Escriba, C., and Campo, E. (2008). A review of smart homespresent state and future challenges. *Computer methods and programs in biomedicine*, 91(1):55–81.

Chatterji, S., Kowal, P., Mathers, C., Naidoo, N., Verdes, E., Smith, J. P., and Suzman, R. (2008). The health of aging populations in china and india. *Health Affairs*, 27(4):1052–1063.

Chen, J., Kam, A. H., Zhang, J., Liu, N., and Shue, L. (2005). Bathroom activity monitoring based on sound. In *International Conference on Pervasive Computing*, pages 47–61. Springer.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13 (1): 2127.*

CUI inc. (2003). CMA-4544PF-W. [Online; accessed 10-Dec-2016].

Fu, Z., Lu, G., Ting, K. M., and Zhang, D. (2011). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303–319.

Goetze, S., Schroder, J., Gerlach, S., Hollosi, D., Appell, J.-E., and Wallhoff, F. (2012). Acoustic monitoring and localization for social care. *Journal of Computing Science and Engineering*, 6(1):40–50.

Guyot, P., Pinquier, J., Valero, X., and Alias, F. (2013). Two-step detection of water sound events for the diagnostic and monitoring of dementia. In *2013 IEEE International Conference on Multimedia and*

Expo (ICME), pages 1–6, San Jose, California (USA). IEEE.

Hollosi, D., Goetze, S., Appell, J., and Wallhoff, F. (2011). Acoustic applications and technologies for ambient assisted living scenarios. In *Proceedings of the AAL Forum*, Lecce (Italy).

Lafortune, G. and Balestat, G. (2007). Trends in severe disability among elderly people. *OECD Health Working Papers*.

Liang, S. and Fan, X. (2014). Audio content classification method research based on two-step strategy. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 5:57–62.

Matern, D., Condurache, A., and Mertins, A. (2013). Adaptive and automated ambiance surveillance and event detection for ambient assisted living. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7318–7321. IEEE.

Melmstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*.

Morsi, Y. S. and Shukla, A. (2015). *Optimizing Assistive Technologies for Aging Populations*. IGI Global.

National Institute on Aging (2007). *Growing Older in America: The Health and Retirement Study*. U.S. Department of Health and Human Services, Washington, DC.

NVIDIA Corp. (2014). NVIDIAs next generation cuda compute architecture: Kepler TM GK110/210. *Available online: (last accessed 20th Oct 2016)*.

NVIDIA Corp. (2016). JETSON TK1. Unlock the power of the GPU for embedded systems applications.

Suzman, R. and Beard, J. (2015). *Global health and aging–Living longer*. National Institute on Aging.

Vacher, M., Portet, F., Fleury, A., and Noury, N. (2010). Challenges in the processing of audio channels for ambient assisted living. In *e-Health Networking Applications and Services (Healthcom), 2010 12th IEEE International Conference on*, pages 330–337. IEEE.

Vacher, M., Portet, F., Fleury, A., and Noury, N. (2013). Development of audio sensing technology for ambient assisted living: Applications and challenges. *Digital Advances in Medicine, E-Health, and Communication Technologies*, page 148.

Valero, X. and Alías, F. (2012). Classification of audio scenes using narrow-band autocorrelation features. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2076–1465. IEEE.

van Hengel, P. and Anemüller, J. (2009). Audio event detection for in-home care. In *Int. Conf. on Acoustics (NAG/DAGA)*.

Wang, J.-C., Lee, H.-P., Wang, J.-F., and Lin, C.-B. (2008). Robust environmental sound recognition for home automation. *IEEE transactions on automation science and engineering*, 5(1):25–31.

Zhang, M.-L. and Zhou, Z.-H. (2005). A k-nearest neighbor based algorithm for multi-label classification. In *2005 IEEE international conference on granular computing*, volume 2, pages 718–721. IEEE.