Audio-based Event Recognition System for Smart Homes

Anastasios Vafeiadis*, Konstantinos Votis*, Dimitrios Giakoumis*, Dimitrios Tzovaras*, Liming Chen† and Raouf Hamzaoui†
*Information Technologies Institute
Center of Research & Technology - Hellas, Thessaloniki, Greece
Email: anasvaf, kvotis, dgiakoum, tzovaras@iti.gr
†Faculty of Technology
De Montfort University
Leicester, UK
Email: liming.chen, rhamzaoui@dmu.ac.uk

***Abstract*—Building an acoustic-based event recognition system for smart homes is a challenging task due to the lack of high-level structures in environmental sounds. In particular, the selection of effective features is still an open problem. We make an important step toward this goal by showing that the combination of Mel-Frequency Cepstral Coefficients, Zero-Crossing Rate, and Discrete Wavelet Transform features can achieve an F1 score of 96.5% and a recognition accuracy of 97.8% with a gradient boosting classifier for ambient sounds recorded in a kitchen environment.**

***Keywords*-Smart homes; assisted living; activity recognition; audio feature extraction; classification; mel-frequency; zero-crossing rate; wavelets;**

I. Introduction

Assisted living in Smart Homes (SH) can change the way millions of elderly people live, manage their conditions and maintain well-being in the future [1]. This could support the ageing population to live longer independently and to enjoy comfort and quality of life in their private environments. While current monitoring and assistance technologies are selectively deployed due to high cost, limited functionality and interoperability issues, future SH could leverage cheap ubiquitous sensors and interconnected smart objects, providing robust context inference and interaction techniques [2]. The next generation of SH technologies will be adaptive to fit versatile living environments, and interoperable for heterogeneous applications. In addition, a service-oriented cloud-based system architecture will support reconfiguration and modular design that is essential to empower care providers to customize solutions.

With the increasing ageing population and the growing demand on novel health care models, research on SH for independent living, self-management and well-being has intensified over the last decade due to the wide availability of affordable sensing and effective processing technologies. Yet, it remains a challenge to develop and deploy SH solutions that can handle everyday life situations and support a wide range of users and care applications. SH technologies must be interoperable for seamless technology integration and rapid application development, and adaptable for easy deployment and management, achieved by thorough testing and validation in multiple application scenarios. This requires a joint multi-disciplinary cross-sector effort of research and development.

In this paper, we concentrate on unobtrusive methods for activity recognition, in particular when using a microphone for Acoustic Event Detection (AED).

Contemporary activity recognition methods in smart homes rely mostly on sensors, which are further separated into wearable [3] and environment-related ones [4]. Recent work [5] shows that ontologies and semantic technologies have been used for activity modeling and representation. Wearable-based techniques depend on user interaction with the sensor and, in most cases, on user motion measured with accelerometers.

Detecting abnormalities in daily home activities necessitates an "always-on" and unobtrusive monitoring system. Gietzelt et al. [6] evaluated gait parameters measured by a single waist mounted accelerometer during everyday life of patients with dementia. Marschollek et al. [7] developed an unobtrusive method to estimate fall risk based on the use of motion sensor data. Palmerini et al. [8] measured the acceleration of the low back to differentiate gait patterns in healthy adults and those with Parkinson's disease (PD). A number of studies [9]–[11] have also taken the first steps to characterize the indoor sound environment and the classification of events.

He et al. [12] provide the time series from such sensors as input to autoregressive models to extract features and classify them utilizing a Support Vector Machine (SVM) classifier. Subsequent works [13], [14] use more advanced models to extract features, namely autoregressive combined with signal magnitude area and tilt angles, while employing modules to further enhance the data separability. Plötz et al. [15] achieve state-of-the-art results with layered Restricted Boltzmann Machines. Considering methodologies focusing on sensors attached to the environment, there is a significant diversity of types, ranging from light sensors, humidity ones,

thermometers and others. Since the extracted information from such sensors can be insufficient, they are usually accompanied by accelerometers [16], which are attached to objects of interest. However, this category of methods fails when activities do not involve the registered objects. Several works [17], [18] place sensors on objects and track their movement, thus activities are inferred through the traces of the objects which are modeled using a Hidden Markov Model (HMM) or a Deep Belief Network (DBN), respectively. Thomas et al. [19] propose a two-step procedure that relies on smart environments. The first step comprises the discretization of activity patterns while the second step trains an Artificial Neural Network (ANN) based on the temporal relations of those patterns. Many features were proposed for Computational Auditory Scene Recognition (CASR). However, the majority work well for structured data, such as speech and non-speech separation or music genre classification. Features in the time domain, such as the Zero-Crossing Rate (ZCR), frequency domain (band-energy ratio, spectral roll-off, spectral flux, spectral centroid, etc.) and in the quefrequency domain, Mel-Frequency cepstral coefficients (MFCCs) are commonly used in the literature [20]–[22].

The fundamental difficulty of non-speech recognition and in particular, environmental sound recognition, is that the input signal is highly variable due to different environmental (indoor, outdoor) and acoustic conditions [23]. In order to extract the features needed, one must use a suitable feature extraction technique.

An AED system involves two phases: training and recognition (testing). During the training phase, a known input signal is recorded and parametric representation of the voice is extracted and stored in the database. During the recognition phase, for a given input signal the features are extracted and the AED system compares it with the reference templates to recognize the utterance.

In general, the modules that are required to develop an auditory recognition system are as follows:

1) Signal acquisition
2) Feature Extraction
3) Acoustic Modeling
4) Recognition

The main problem that this paper tries to address is feature and classifier selection for our specific application, i.e., indoor audio-based human activity recognition. We show that the use of hybrid features (ZCR, MFCC, and Discrete Wavelet Transform (DWT) coefficients) with a gradient boosting classifier gives high recognition accuracy. To the best of our knowledge, this approach had not been used in the context of environmental sound classification in an SH.

This paper is organized as follows. Section II gives an overview of the proposed system (signal acquisition, feature extraction, classification). Section III presents the experimental setup. Section IV gives the results. Finally, Section V concludes the paper.

II. Proposed System

In order to avoid determining the range of frequencies that are relevant to identifying the kitchen environmental sound, we had to split the input signal into smaller frames for processing. No information was lost using this approach. In the time domain we calculated the ZCR features of the signal and in the frequency domain the MFCCs (static, first and second order derivatives). The Discrete Wavelet Transform (DWT) provided useful features in both the time and frequency domain. Well-known classifiers, such as, k-nearest neighbor, SVM, Random Forest, Extra Trees and Gradient Boosting were used for classification and performance evaluation (Fig.1). The details of the process of feature extraction and classification are given in the following sub-sections.

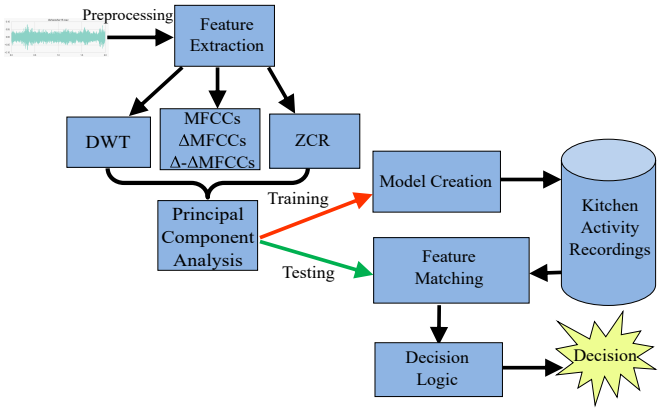

Figure 1. Proposed AED system

A. Signal acquisition

The success of the signal recording depends on the recording environment and the placement of the microphone. Ideally, the recordings should take place in soundproof studios or labs. However, this is not possible in real life. Therefore, we had to examine test case scenarios with various types of noises that could occur in a noisy environment.

In this first step of the preprocessing, we recorded the input signal in mono using 44,100 Hz as the sampling frequency. This allowed us to use frequencies up to 22,050 Hz, satisfying the Nyquist criteria. This maximum frequency is sufficient to cover all the harmonics generated by our input signal and removes noise above this range (also not detected by human ear).

B. Feature Extraction

An AED system relies on feature extraction from the input signal. The aim of feature extraction is to reduce the amount of data present in a given sound signal while retaining

the discriminative information of the source [24]. Feature extraction plays a crucial role in the overall performance of the system. There are a great many feature extraction techniques, including:

- Linear Predictive Analysis (LPC)
- Zero-Crossing Rate (ZCR)
- Linear Predictive Cepstral Coefficients (LPCC)
- Perceptual Linear Predictive coefficients (PLP)
- Mel-Frequency Cepstral Coefficients (MFCC)
- Power spectral analysis (FFT)
- Mel scale Cepstral analysis (MEL)
- Relative spectra filtering of the log domain coefficients (RASTA)
- First order derivative (DELTA)
- Second order derivative (DELTA - DELTA)
- Discrete Wavelet Transform (DWT)

In our work, we used a hybrid approach, combining MFCC, ZCR and wavelet features. MFCC and ZCR features are well-known and widely used in speech recognition. However, their performance was severely affected by the high levels of noise present within a home environment. For this reason, we added wavelet features to improve system robustness to noise.

1) MFCC: Mel-Frequency Cepstral Coefficients: Mel-Frequency Cepstral Coefficients (MFCC) is one of the most commonly used feature extraction techniques used in voice recognition [25]. The use of Mel-Frequency Cepstral Coefficients can be considered one of the standard methods for feature extraction [26]. The use of 12 cepstral (we excluded the 0^{th} coefficient; DC component) coefficients proved to be ideal for feature extraction, that could be used as inputs to a classifier. Fig.2 shows the steps involved in MFCC feature extraction.

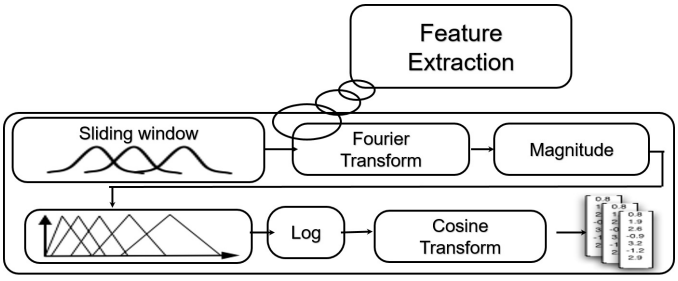

Figure 2. MFCC Feature Extraction

The input signal must first be broken up into small sections, each of N samples. In order to avoid loss of information, 50% frame overlap is used. Each frame begins at some offset of K samples with respect to the previous frame where $K \leq N$.

For each frame, a windowing function is usually applied to increase the continuity between adjacent frames. Common windowing functions include the rectangular window, the Hamming window, the Blackman window and flattop window. We have used the Hamming window as it is the most commonly used window function in audio signal processing [27].

The Discrete Fourier Transform (DFT) turns the windowed sound segment into the frequency domain and the short-term power spectrum *P(f)* is obtained.

The spectrum *P(f)* is warped along its frequency axis f (in Hertz) into the mel-frequency axis as *P(M)*, where *M* is the mel-frequency using Eq.(1). This is to approximately reflect the human ear perception

$$M(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{1}$$

The resulted warped power spectrum is convolved with the triangular band-pass filter *P(M)* into $\theta(M)$. The process of convolution with the relatively broad critical-band masking curves $\psi(M)$ significantly reduces the spectral resolution of $\theta(M)$ in comparison with the original *P(f)*, which allows the down sampling of $\theta(M)$. The discrete condition of $\psi(M)$ with $\theta(M)$ yields samples of the critical-band power spectrum as $\theta(M_{\mathrm{k}})$, *k = 1,...,K* in Eq.(2) where $k's$ are linearly spaced in the mel-frequency scale. Afterwards, *K* outputs *X(k)* = $\ln \theta(M_{\mathrm{k}})$), where *k = 1,...,K* are obtained. When implemented, $\theta(M_{\mathrm{k}})$ is the average of the samples of the power spectrum rather than the sum of all samples.

$$\theta(M_{\mathrm{k}}) = \sum_{M} P(M - M_{\mathrm{k}})\psi(M), \quad k = 1, \ldots, K \tag{2}$$

The static MFCCs are computed using Eq.(3)

$$c_{\mathrm{n}} = \sum_{k=1}^{K} \cos[n(k - 0.5)\tfrac{\pi}{K}], \quad n = 1, \ldots, K \tag{3}$$

Advantage: As the frequency bands are positioned logarithmically in MFCC, the human system response is approximated more closely than any other system.

Disadvantage: MFCC values are not very robust in the presence of additive noise, and so it is common to normalize their values in speech recognition systems to lessen the influence of noise.

As mentioned, MFCCs are used for voice/speaker recognition. However, in our case the signals had significant information at the trajectories of the MFCC coefficients over time Therefore, we had to calculate the delta and delta-deltas, also known as differential and acceleration coefficients. The delta coefficients are calculated using Eq.(4)

$$\Delta_c[m] = \frac{\sum_{i=1}^{K} i(c[m+i] - c[m-i])}{2\sum_{i=1}^{K} i^2} \tag{4}$$

where $\Delta_c[m]$ is the differential coefficient, from a frame m computed in terms of the static MFCC coefficients $c[m+i]$ to $c[m-i]$ and i denotes the frame number corresponding

to the time-domain frame. The acceleration coefficients are calculated similarly from the deltas.

For our approach, we had 12 cepstral coefficients + 1 energy coefficient, 12 delta cepstral coefficients + 1 delta energy coefficient and 12 double delta cepstral coefficients + 1 double delta energy coefficient; making a total of 39 MFCC features.

2) DWT: Discrete Wavelet Transform: Any environmental signal is a non-stationary signal. The Fourier Transform (FT) is not suitable for the analysis of such non-stationary signal because it provides only the frequency information of signal but does not provide information about the time at which the specific frequency is present. The windowed short-time FT (STFT) provides the temporal information about the frequency content of signal. A drawback of the STFT is its fixed time resolution due to fixed window length [28].

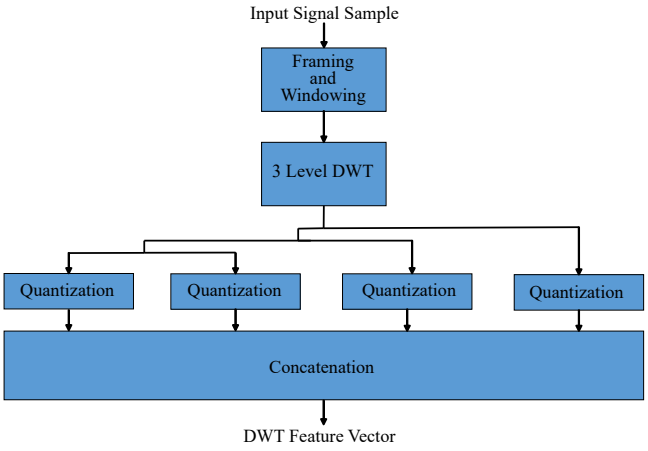

Figure 3. DWT pyramidal algorithm

The discrete wavelet transform (DWT) provides a compact representation of the signal in time and frequency that can be computed efficiently using a fast, pyramidal algorithm (Fig.3) related to multirate filter banks. As a multirate filter bank, DWT can be viewed as a constant Q filter bank. Each subband contains half the samples of the neighboring higher frequency subband. In the pyramidal algorithm the input signal is analyzed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information. The coarse approximation is then further decomposed using the same wavelet decomposition step. This is achieved by successive high pass and low pass filtering of the time domain signal. In our implementation, we used a three-level DWT decomposition with the db20 wavelet [29], since they proved to be more robust to noise. The wavelet transform concentrated the signal features in a few large-magnitude wavelet coefficients; hence the coefficients with a small value (noise) could be removed without affecting the input signal quality.

In the kitchen environment signals, high frequency components are present very briefly at the onset of a sound while lower frequencies are present for a long period. DWT resolves all these frequencies in a very satisfactory manner. The DWT parameters contain the information of different frequency scales. This helps in getting the input signal information of the corresponding frequency band.

3) ZCR: Zero-Crossing Rate: In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. This average zero-crossing rate gives a reasonable way to estimate the frequency of sine wave.

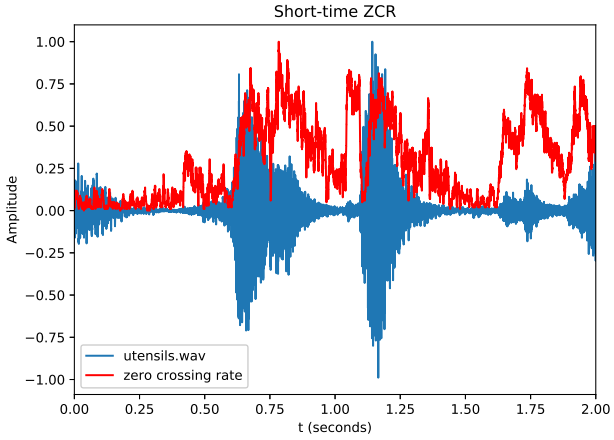

Figure 4. Short-time ZCR calculation of the sound signal of kitchen utensils (forks and spoons)

Environmental signals are broadband signals and interpretation of average zero-crossing rate is therefore much less precise [30]. However, rough estimates of spectral properties can be obtained using a representation based on the short-time average zero-crossing rate as shown in Fig.4.

In this implementation, the zero-crossing rate was calculated for each 20 ms frame of a sample's data. Then the local variance of the ZCR was calculated over each second of data (50 frames per data second). Finally, the mean of the local variances was taken to be the sample's data value for the ZCR variance feature.

4) PCA: Principal Component Analysis : The central idea of principal component analysis (PCA) is to reduce the dimensionality of a dataset that consists of many interrelated variables, while retaining as much as possible of the variation present in the dataset [31]. This is done by projecting the original feature vector onto principal component axes [32]. These axes are orthogonal and correspond to the directions of the greatest variance in the original feature space. Projecting input vectors onto the principal subspace helps reducing the redundancy in original feature space and

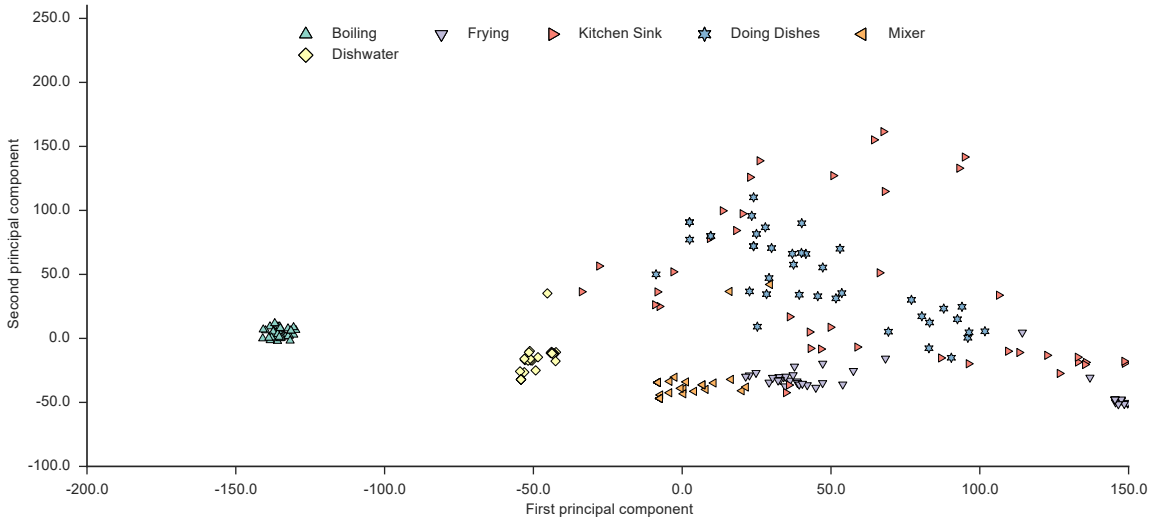

Figure 5. Principal Component Analysis

dimension as well. In this work, we have applied the PCA technique to MFCC, ZCR and wavelet features to extract the most significant components. For each five-second training recording, a total of 8250 features were used as the input for the classifier. With PCA, we reduced the feature space down to two principal components (Fig.5). The new features obtained from PCA were used as inputs to the classifier.

C. Feature Classification

The classification step in automatic signal identification systems is in fact a feature matching process between the features of a new input signal and the features saved in the database. For our experiment we have compared the performance of a kNN classifier with 5 nearest neighbors, an SVM with a linear and a Radial Basis Function (RBF) kernel, an Extra Trees classifier, a Random Forest and finally the Gradient Boosting classifier.

III. Experimental Setup

For our experiments, we used a mobile phone to record sounds of activities in the kitchen as described in Fig.6. In addition, we used similar sounds from Freesound [33]. To check the robustness of our AED system, we masked some of the recordings with noise (e.g. sounds from other devices, speech, synthetic noise). A total of 1080 different signals from different activities were collected (180 kitchen faucet, 180 boiling, 180 frying, 180 dishwasher, 180 mixer, 180 doing dishes). The setup included the following steps:

- 180 audio signals for each class were collected using the mobile application and the Freesound repository. The mobile application started recording automatically every 20 seconds for 5 seconds

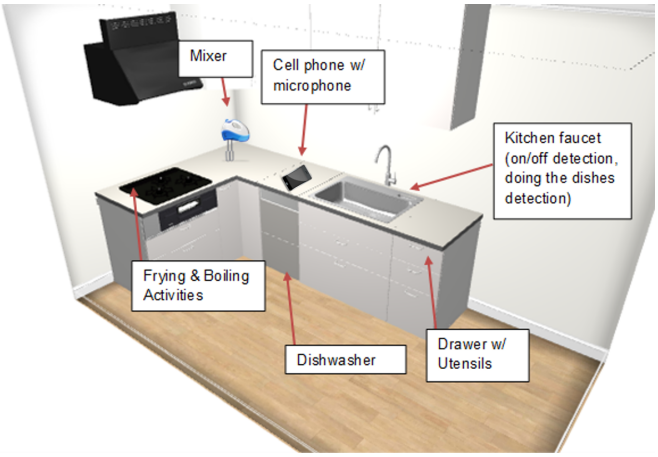

Figure 6. Kitchen Environment Setup

- from the dataset that was created, the MFCCs (static, first and second derivatives), ZCR and DWT features for each signal were extracted and a new dataset was formed
- Monte Carlo cross-validation was used to randomly split the dataset into training and validation (testing) data and the results were averaged over the splits
- we extracted the MFCC, ZCR and DWT features of the validation data
- PCA was applied for dimensionality reduction
- different classifiers were used for activity recognition

The implementation of the server that handles the feature extraction and classification was based on Flask RESTful API service for Python. Hence, the server responded with:

- name of the test file

Table I
CLASSIFIER PERFORMANCE COMPARISON

MFCC+DWT+ZCR Extraction				
Classifier	**PRECISION**	**RECALL**	**F1-SCORE**	**ACCURACY**
kNN (5 nearest neighbors)	90.3%	95.3%	92.8%	94.3%
SVM (linear kernel)	96.2%	90.4%	93.2%	96.7%
SVM (RBF kernel)	97.4%	90.2%	93.7%	97.2%
Extra Trees	95.7%	89.4%	92.4%	96.1%
Random Forest	94.0%	92.2%	93.1%	95.9%
Gradient Boosting	***96.3%***	***96.7%***	***96.5%***	***97.8%***

- classification result

IV. RESULTS

We focused on the extraction of the appropriate features for classification. For instance, by using only the DWT coefficients, we compared the results for a Haar "mother" wavelet and the Daubechies20 and we achieved an accuracy of 84.2% and 89.3% respectively. For this test, we used the wavelet features and an SVM classifier with a linear kernel.

However, introducing the MFCCs as well as the ZCR features improved the performance of the AED system. Extracting the selected features from the five-second signals and comparing the performance of different classifiers, we obtained the results shown in Table I.

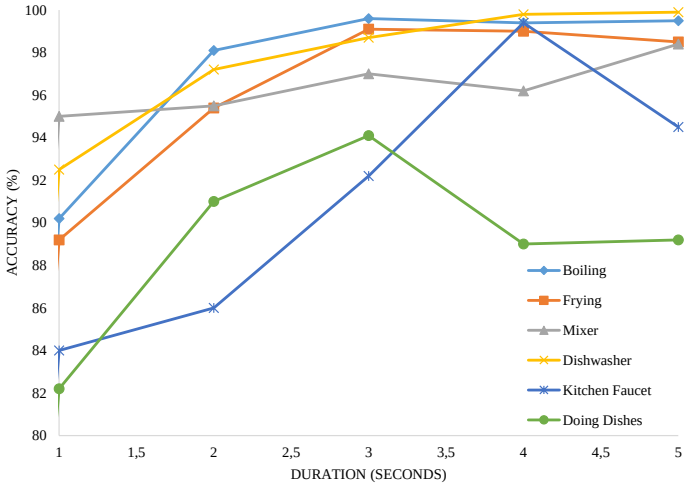

Figure 7. Recognition accuracy (using the Gradient Boosting classifier) as a function of the sample duration

For the Random Forest classifier, we noticed, as the theory suggests, that a higher number of trees can give better performance, with a smaller risk of overfitting. The number of leaves in the tree was set to 50. We selected a small number to capture noisy instances in the training dataset. For the SVM classifier, the best results were achieved with an RBF kernel where $\sigma = 1$ and $C = 0.1$. The parameter σ of the RBF kernel handles the non-linear classification. It is a similarity measure between two points. C is the cost of classification. Finally, for the Gradient Boosting we picked 500 estimators. Gradient boosting is fairly robust to overfitting, therefore this large number resulted in a better performance, achieving an F1-Score of 96.5%. We obtained solid results for boiling, frying, the use of mixer, and also the use of dishwasher. However, we noticed that the activity of the "running" kitchen faucet was understood by our system as doing the dishes. This is because some recordings were very similar, due to the timing (meaning that no dishes or utensils were "heard" from the microphone). The Gradient Boosting classifier gave us the best results since we noticed a stable relation between the precision, recall, F1-Score and recognition accuracy.

Furthermore, we studied the impact of segment duration on the accuracy of activity recognition within the kitchen environment.

Table II
COMPARISON OF RECOGNITION ACCURACY BETWEEN THE PROPOSED SYSTEM AND A BASELINE SYSTEM BASED ON MFCC AND GMM FOR THE DCASE 2016 DATASET [34]

Class	Baseline MFCC GMM (%)	MFCC+ZCR+Wavelets Gradient Boosting (%)
Beach	69.3	90.2
Bus	79.6	91.0
Cafe/Restaurant	83.2	80.6
Car	87.2	91.4
City center	85.5	82.1
Forest path	81.0	95.3
Grocery store	65.0	89.9
Home	82.1	90.0
Library	50.4	65.1
Metro station	94.7	88.4
Office	98.6	95.6
Park	13.9	60.6
Residential area	77.7	71.3
Train	33.6	61.2
Tram	85.4	88.6

Fig. 7 shows that a three-second time duration of the input signal is sufficient for accurate activity recognition.

However, we noticed an unexpected drop-off for the activity of doing the dishes after the third second. Examination of the confusion matrices revealed that there is a confusion

between the activity of doing the dishes and the operation of the kitchen sink. After careful listening of all the recordings, we noticed that there were times that the user had the faucet open and only at the last second of the recording he/she picked an object (plate, utensils) to wash.

To further validate our approach, we tested it on a DCASE 2016 dataset. DCASE 2016 [35] is the latest IEEE Audio and Acoustic Signal Processing (AASP) challenge for scene classification and polyphonic event detection. We compared our results to those of the DCASE2016 baseline system, which uses MFCCs for feature extraction and a Gaussian Mixture Model (GMM) as a classifier [34]. Our approach was able to achieve higher accuracies for most of the given classes. Table II shows that for some classes, the two approaches have some strengths and weaknesses. This is due to the different datasets. Our approach works well for indoor sounds, especially within a home, and not so well for some outdoor scenes where there is a high presence of noise.

V. Conclusions

We presented a system that is able to use environmental sounds for event detection in a smart home. While our experiments were done in a kitchen environment, our approach is flexible enough to be applied to other smart home environments. One potential application of our system is in real-time unobtrusive ambient assisted living.

Although we obtained excellent results, it must be noted that the problem of audio-based event recognition remains a hard task. This is because features and classifiers that work extremely well for a specific dataset may fail for another.

As future work, we plan to consider other features and other classifiers (e.g., GMM, HMM, Convolutional Neural Networks) to improve the recognition accuracy of our system, including in the challenging case of polyphonic event detection.

VI. Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676157, project ACROSSING.

References

[1] J. Nehmer, M. Becker, A. Karshmer, and R. Lamm, "Living assistance systems: An ambient intelligence approach," in *Proceedings of the 28th International Conference on Software Engineering*, ser. ICSE '06. New York, NY, USA: ACM, 2006, pp. 43–50.

[2] N.-C. Chi and G. Demiris, "A systematic review of telehealth tools and interventions to support family caregivers," *Journal of Telemedicine and Telecare*, vol. 21, no. 1, pp. 37–44, 2015.

[3] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790–808, Nov. 2012.

[4] 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, Mar. 2013.

[5] L. Chen, C. Nugent, and G. Okeyo, "An ontology-based hybrid approach to activity modeling for smart homes," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 92–105, Feb. 2014.

[6] M. Gietzelt, K. H. Wolf, M. Kohlmann, M. Marschollek, and R. Haux, "Measurement of accelerometrybased gait parameters in people with and without dementia in the field a technical feasibility study," *Methods of Information in Medicine*, vol. 52, no. 4, pp. 319–325, 2013.

[7] M. Marschollek, A. Rehwald, K. H. Wolf, M. Gietzelt, G. Nemitz, H. Meyer zu Schwabedissen, and R. Haux, "Sensor-based fall risk assessment - an expert 'to go'," *Methods of Information in Medicine*, vol. 50, no. 5, pp. 420–426, 2011.

[8] L. Palmerini, S. Mellone, G. Avanzolini, F. Valzania, and L. Chiari, "Quantification of motor impairment in parkinson's disease using an instrumented timed up and go test," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 4, pp. 664–673, 2013.

[9] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, *Acoustic Event Detection and Classification*. London: Springer London, 2009, pp. 61–73.

[10] H. Lozano, I. Hernáez, A. Picón, J. Camarena, and E. Navas, *Audio Classification Techniques in Home Environments for Elderly/Dependant People*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 320–323.

[11] R. M. Alsina-Pags, J. Navarro, F. Alas, and M. Hervs, "homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring," *Sensors*, vol. 17, no. 4, 2017.

[12] Z.-Y. He and L.-W. Jin, "Activity recognition from acceleration data using ar model representation and svm," in *2008 International Conference on Machine Learning and Cybernetics*, vol. 4, 2008, pp. 2245–2250.

[13] A. M. Khan, Y.-K. Lee, S. Lee, and T.-S. Kim, "Accelerometer's position independent physical activity recognition system for long-term activity monitoring in the elderly," *Med. Biol. Engineering and Computing*, vol. 48, pp. 1271–1279, 2010.

[14] A. M. Khan, Y. K. Lee, S. Y. Lee, and T. S. Kim, "A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 5, pp. 1166–1172, 2010.

[15] T. Plötz, N. Y. Hammerla, and P. Olivier, "Feature learning for activity recognition in ubiquitous computing," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Two*, ser. IJCAI'11. AAAI Press, 2011, pp. 1729–1734.

[16] J. Yin, Q. Yang, and J. J. Pan, "Sensor-based abnormal human-activity detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 8, pp. 1082–1090, Aug. 2008.

[17] M. Buettner, R. Prasad, M. Philipose, and D. Wetherall, "Recognizing daily activities with rfid-based sensors," in *Proceedings of the 11th International Conference on Ubiquitous Computing*, ser. UbiComp '09. New York, NY, USA: ACM, 2009, pp. 51–60.

[18] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *Proceedings of the 11th IEEE International Conference on Computer Vision*, Oct. 2007, pp. 1–8.

[19] S. T. M. Bourobou and Y. Yoo, "User activity recognition in smart homes using pattern clustering applied to temporal ann algorithm," *Sensors*, vol. 15, no. 5, pp. 11 953–11 971, 2015.

[20] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, May 2002, pp. II–1941–II–1944.

[21] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[22] M. Perttunen, M. V. Kleek, O. Lassila, and J. Riekki, "Auditory context recognition using svms," in *Proceedings of the 2nd International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, 2008, pp. 102–108.

[23] M. Cernak, A. Lazaridis, A. Asaei, and P. N. Garner, "Composition of deep and spiking neural networks for very low bit rate speech coding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2301–2312, Dec. 2016.

[24] M. Myllymki and T. Virtanen, "Voice activity detection in the presence of breathing noise using neural network and hidden markov model," in *2008 16th European Signal Processing Conference*, Aug. 2008, pp. 1–5.

[25] B. Milner, J. Darch, I. Almajai, and S. Vaseghi, "Comparing noise compensation methods for robust prediction of acoustic speech features from mfcc vectors in noise," in *2008 16th European Signal Processing Conference*, Aug. 2008, pp. 1–5.

[26] M. Boussaa, I. Atouf, M. Atibi, and A. Bennis, "Comparison of mfcc and dwt features extractors applied to pcg classification," in *2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA)*, Oct. 2016, pp. 1–5.

[27] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[28] H. Varshney, M. Hasan, and S. Jain, "Energy efficient novel architectures for the lifting-based discrete wavelet transform," *IET Image Processing*, vol. 1, no. 3, pp. 305–310, 2007.

[29] S. Bilgin, O. Polat, and O. H. Colak, "The impact of daubechies wavelet performances on ventricular tachyarrhythmia patients for determination of dominant frequency bands in hrv," in *2009 14th National Biomedical Engineering Meeting*, May 2009, pp. 1–4.

[30] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, *Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy*. Dordrecht: Springer Netherlands, 2010, pp. 279–282.

[31] I. T. Jolliffe, *Principal component analysis*, ser. Springer series in statistics. New York, Berlin, Heidelberg: Springer, 2002.

[32] N. Kaberpanthi and A. Datar, "Article: Speaker independent speech recognition using mfcc with cubic-log compression and vq analysis," *International Journal of Computer Applications*, vol. 95, no. 26, pp. 33–37, 2014.

[33] Robinhood76. (2008) Kitchen common sounds. [Online]. Available: https://www.freesound.org/people/Robinhood76/packs/3870

[34] T. Heittola, A. Mesaros, and T. Virtanen, "DCASE2016 baseline system," DCASE2016 Challenge, Tech. Rep., 2016.

[35] (2016) Dcase2016 challenge. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/challenge