Audio Content Analysis for Unobtrusive Event Detection in Smart Homes

Anastasios Vafeiadis[a,1,*], Konstantinos Votis[a], Dimitrios Giakoumis[a], Dimitrios Tzovaras[a], Liming Chen[b], Raouf Hamzaoui[b]

[a] *Center for Research and Technology Hellas - Information Technologies Institute, Thessaloniki, Greece*
[b] *Faculty of Computing, Engineering and Media, De Montfort University, Leicester, UK*

Abstract

Environmental sound signals are multi-source, heterogeneous, and varying in time. Many systems have been proposed to process such signals for event detection in ambient assisted living applications. Typically, these systems use feature extraction, selection, and classification. However, despite major advances, several important questions remain unanswered, especially in real-world settings. This paper contributes to the body of knowledge in the field by addressing the following problems for ambient sounds recorded in various real-world kitchen environments: (1) which features and which classifiers are most suitable in the presence of background noise? (2) what is the effect of signal duration on recognition accuracy? (3) how do the signal-to-noise-ratio and the distance between the microphone and the audio source affect the recognition accuracy in an environment in which the system was not trained? We show that for systems that use traditional classifiers, it is beneficial to combine gammatone frequency cepstral coefficients and discrete wavelet transform coefficients and to use a gradient boosting classifier. For systems based on deep learning, we consider 1D and 2D Convolutional Neural Networks (CNN) using mel-spectrogram energies and mel-spectrograms images as inputs, respectively, and show that the 2D CNN outperforms the 1D CNN. We obtained competitive classification results

*Corresponding author
Email address: anasvaf@iti.gr (Anastasios Vafeiadis)

for two such systems. The first one, which uses a gradient boosting classifier, achieved an F1-Score of 90.2% and a recognition accuracy of 91.7%. The second one, which uses a 2D CNN with mel-spectrogram images, achieved an F1-Score of 92.7% and a recognition accuracy of 96%.

Keywords: Smart homes; ambient assisted living; audio signal processing; feature extraction; feature selection; deep learning

1. Introduction

Smart home-based ambient assisted living Information and Communications Technology (ICT) solutions can allow the elderly to remain in their own homes and live independently for longer [1]. Research on ICT solutions for ambient assisted living has intensified over the last decades considerably, due to the emergence of affordable powerful sensors and progress in artificial intelligence [2, 3, 4].

Various human activity recognition (HAR) systems that monitor daily activities to identify abnormal behavior have been proposed for ambient assisted living applications [5, 6].

One common approach to automated HAR uses portable sensors such as accelerometers and gyroscopes [7, 8]. However, these sensors require cooperation of the subject, may restrict body movement, and are energy constrained [9, 10]. Another approach relies on computer vision [11, 12]. However, privacy concerns are hindering its adoption. A further approach is based on audio processing. Features are extracted from the environmental sounds and classifiers are used to recognize the corresponding human activity [13, 14, 15].

While several audio-based HAR systems have been proposed, a number of important questions remain unanswered:

- which features and which classifiers are most suitable in the presence of background noise?
- what is the effect of the duration of the signal segment used for classification on recognition accuracy? Decreasing the segment duration decreases

the response time of the system but may harm its recognition accuracy. At the same time, increased duration can lead to increased co-occurrence of multiple events within the same sound segment;

- how do the signal-to-noise-ratio (SNR) and the distance between the microphone and audio source affect the recognition accuracy in a new environment (i.e., one which was not used to train the classifier)?

Our work answers these questions for a real-world indoor kitchen environment where large audio datasets are captured and processed to train classifiers. Two representative acoustic event detection (AED) approaches are studied. The first one extracts time and frequency features and uses a traditional classifier. We compared various features and classifiers and showed that the best results are obtained with hybrid time-frequency features, together with a gradient boosting classifier. Our best system achieved an F1-score of 90.2% and a recognition accuracy of 91.7%. The second system uses mel-spectrogram images of the audio signals as input to a 2D CNN. We showed that compared to a 1D CNN that applies max-pooling to only one dimension, applying max pooling to both dimensions of the input (time and frequency) reduces the dimensionality in a more uniform manner, yielding more salient features with each consecutive convolutional-max pooling operation. This approach achieved a recognition accuracy of 96% and an F1-Score of 92.7%. Additionally, we observed that in a real-world environment the recognition accuracy for some classes did not improve when the signal duration was greater than 3 s. This was due to overlapping sounds that occurred in the kitchen environment (e.g., kitchen faucet running, while the user picks a plate to wash). Even in the cases where the recognition accuracy increased, the improvement was not significant. Studying the trade-off between signal duration and accuracy is important in scenarios where the data needs to be captured and processed on a system-on-chip device (e.g., Raspberry Pi), with limited memory size. Finally, since real-world environments typically include noise, we studied the effect of the SNR and distance between the microphone and the target audio event on the recognition accuracy. For events such

as using the mixer and the utensils (forks, spoons, knives), the recognition accuracy was high despite the background noise of a kitchen fan and a refrigerator. The high amplitudes in the signals associated with these events could mask the low amplitudes of the background noise signals. On the other hand, we noticed a drop in the recognition accuracy for quieter sounds (e.g., dishwasher). We did not add artificial background noise to affect the SNR since we wanted to be as close to a real-world scenario as possible. The classification results that we obtained at various distances showed that we can achieve good accuracies with one microphone. This was useful, especially for monitoring houses of the elderly, where the number of sensors should be as small as possible.

The two systems are unobtrusive and preserve privacy as the raw audio is immediately deleted after feature extraction and cannot be recovered from the features.

The rest of the Chapter is organized as follows. Section 3.2 discusses related work. Section 3.3 describes the two systems used in our study, giving details on signal acquisition, feature extraction, feature selection and classification. The experimental setup and the results are presented in Sections 3.4 and 3.5, respectively. Finally, Section 3.6 concludes this Chapter.

2. Related Work

Audio-based activity recognition has received a lot of attention from researchers in recent years [16, 17]. A number of studies have also taken the first steps to characterize the indoor sound environment and the classification of events [18, 19].

While many approaches addressed the problem of audio-based activity recognition in a home environment [20, 21, 22, 23], there is not enough justification for the classifier and feature selection. Most of them used well-known features from the field of speech recognition (e.g., Mel-frequency Cepstral Coefficients (MFCCs)) along with classifiers, such as the k-Nearest Neighbors (kNN) algorithm, to serve as a proof of concept for indoor audio-based activity classifi-

cation. Chu et al. [24] showed that increasing the number of audio features does not improve the recognition accuracy of a system classifying environmental sounds and used the matching pursuit algorithm to obtain effective time-frequency features.

Deep Neural Networks (DNNs) are able to extract important information from the raw data without the need for hand-crafted feature extraction and outperform traditional classifiers in many tasks. There is significant research on recognizing single events in monophonic recordings [25] and multiple concurrent events in polyphonic recordings [26]. Different feature extraction techniques [27], hybrid classifiers [28, 29] and very deep neural models [30] have been explored. However, none of these works compared 1D and 2D CNN architectures for ambient sounds.

Another focus of this work is the duration of the signal used with an audio-based event detection system. The works [31, 32, 33] examined the length needed for sufficient recognition accuracy. They used systems based on time-frequency features and simple classifiers, such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs). The proposed approaches work well with datasets that contain indoor or outdoor environmental sounds. However, due to the high variability in the class and the similarity between different classes, they can fail in a specific AED task (e.g., in a kitchen environment).

Finally, there has been extensive research on the effect of the SNR in the presence of background noise [34, 35, 36]. Wang et al. [37] performed experiments for various artificially added SNRs (0-10 dB and clean recordings) using different environmental sound datasets and a hierarchical-diving deep belief network. However, all previous work assumed prior knowledge of the SNR, which is not possible in a real-world environment.

3. Proposed System Architectures

We propose two approaches for acoustic event detection in an indoor environment.

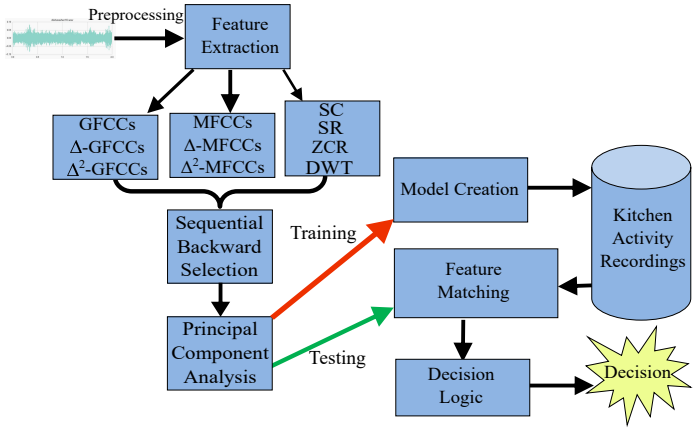


Figure 1: First proposed AED approach.

For the first AED approach (Figure 1), we considered time-domain features (Zero-Crossing Rate (ZCR)), frequency-domain features (MFCCs, Gammatone Frequency Cepstral Coefficients (GFCCs), Spectral Roll-Off (SR), Spectral Centroid (SC)), and time-frequency features (Discrete Wavelet Transform coefficients). Furthermore, we studied the effect of adding many audio features along with proper feature selection and reduction techniques on recognition accuracy. For classification, we examined well-known classifiers such as kNN, SVM, Random Forest, Extra Trees and Gradient Boosting.

For the second AED approach, we used a CNN trained on mel-spectrogram images (Figure 5). We show that even for a small dataset, a 2-dimensional CNN with 2-dimensional max-pooling (downsamping) layers can provide good recognition accuracy results. The details of the two approaches are given in the following sub-sections.

3.1. Signal Acquisition

The success of the signal recording depends on the 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 examined test case scenarios with various types of noises that may occur in a home environment. Three kitchen environments (first author's house, CERTH

KRIPIS smart home and AKTIOS S.A. Elderly Care Units in Vari, Athens) were used for data collection.

In the first step of the preprocessing, we recorded the input signal in stereo at 44,100 Hz (16-bit depth) and then averaged the two channels. This allowed us to use frequencies up to 22,050 Hz, according to the Nyquist criterion. This is sufficient to cover all the harmonics generated by our input signal and removes noise above this range (also not detected by human ear).

3.2. Data Augmentation

Environmental audio recordings have various temporal properties. Therefore, we need to make sure that we have captured all the significant information of the signal in both the time and frequency domain. Any environmental signal is a non-stationary signal [24], since it is a stochastic signal and a signal value is not equally probable to occur given another signal value at any time instance.

Previous research [38, 39] showed that data augmentation can significantly improve the performance of a classification system by introducing variability into the original recordings. For this reason, for both AED approaches, we produced two additional recordings from the original ones. First, for each recording, we added noise with uniform probability distribution. This allowed us to train our system better, since the test audio data in an unknown environment (not used for training) would also include various noises (e.g., different people speaking while performing an activity such as cooking). Second, we re-sampled the original recording from 44.1 kHz to 16 kHz. Most of the monitored kitchen environment recordings (mixer, dishwasher, faucet, utensils) had a fundamental frequency of around 600-700 Hz. We focused on the harmonics produced by devices such as the mixer and the dishwasher and found that a lot of information at around 11 kHz was necessary for these classes.

The quality of the data was maintained since i) downsampling removed the frequencies above 16 kHz and did not affect the general recording since the energy of the highest frequencies (above 16 kHz) was very small and ii) the added uniform noise corresponded to the scenario where ambient noise was

present in the kitchen environment (e.g., fan and refrigerator of the setup in Figure 6 (c)).

3.3. Feature Extraction

To include the range of frequencies that are relevant to identifying the kitchen environmental sounds and to efficiently extract the audio features, we split the input signal into smaller frames for processing. Each frame had a window size of 20 ms with a 10 ms hop size from the next one (50% overlapping sliding Hamming window). Thus, there were 173 frames per recording.

For the second AED approach, we calculated the mel-spectrogram with 128 bins to keep the spectral characteristics of the audio signal while greatly reducing the feature dimension. We normalized the values before using them as an input into the CNN by subtracting the mean and dividing by the standard deviation.

In the following, we give the details of feature extraction for the first AED approach.

3.3.1. MFCC: Mel-Frequency Cepstral Coefficients

MFCCs are one of the most popular features for voice recognition [40]. Figure 2 shows the steps involved in MFCC feature extraction.

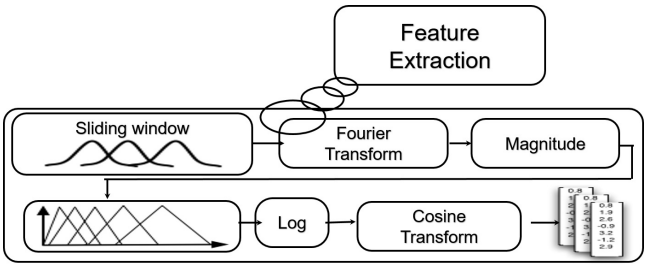


Figure 2: MFCC Feature Extraction

One of the disadvantages of MFCCs is that they are not very robust against additive noise, and so it is common to normalize their values in speech recognition systems to lessen the influence of noise.

MFCCs are used for voice/speaker recognition. However, the indoor environmental audio signals had significant information at the trajectories of the MFCCs over time. Therefore, we included the delta values and delta-delta values [41]. To compute these features, we used the *mfcc* function of the Librosa library [42], which return 39 MFCC features per frame: 13 MFCCs where the zeroth coefficient was replaced with the logarithm of the total frame energy, 13 delta features and 13 delta-delta features.

3.3.2. DWT: Discrete Wavelet Transform

The DWT provides a compact representation of a signal in time and frequency and can be computed efficiently using a fast, pyramidal algorithm. In the pyramidal algorithm the input signal is analyzed at different frequency bands with different resolution by decomposing it into a coarse approximation and detail information. This is achieved by successive high pass and low pass filtering of the time domain signal. We used an 8-level DWT with the 20-coefficient wavelet family (db20) proposed by Daubechies [43], because of its robustness to noise, and extracted the mean and variance in each sub-band, resulting in 16 (high-frequency) features. 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.

3.3.3. 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.

The zero-crossing rate returned a 1×173 vector for each recording and we calculated the mean and median of each vector, resulting in two ZCR features per recording.

3.3.4. SR: Spectral Roll-off

Spectral Roll-off (SR) is defined as the frequency below which a certain percentage (85% - 95%; depending on the application) of the magnitude distribution of the power spectrum is accumulated. The equation of the feature is given in Equation (1):

$$\sum_{k=1}^{m} X_i(k) = C \sum_{k=1}^{N} X_i(k) \tag{1}$$

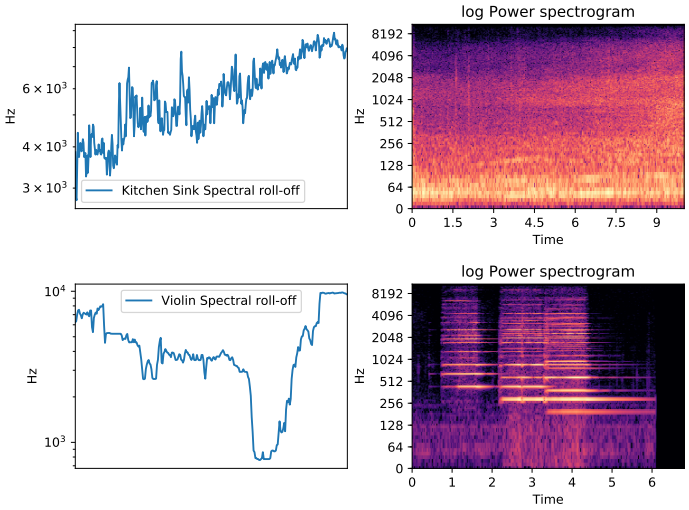


Figure 3: SR comparison between the sound of the kitchen sink (top) and the sound of a violin (bottom). The x-axis on shows time in s

where $X_i(k)$, $k = 1, ..., N$ are the Discrete Fourier Transform (DFT) coefficients of the i-th short-term frame and N is the number of frequency bins. The DFT coefficient $X_i(m)$ corresponds to the SR of the i-th frame and C is the percentage of the magnitude distribution of the spectrum. We found a threshold of 95% to be suitable for distinguishing different kitchen sounds. The mean and median of the SR for each recording were calculated and normalized between 0 and 1.

Figure 3 shows the difference of the SR between a violin recording and the running tap water in the sink. The harmonics of the violin are very distinct in

the spectrum, the mean is 0.423 and the median is 0.417. On the other hand, the mean and median of the kitchen sink sound are 0.811 and 0.803 respectively.

3.3.5. SC: Spectral Centroid

Spectral Centroid (SC) is defined as the "center of gravity" of the spectrum. It is described by Equation (2)

$$SC = \frac{\sum_{k=1}^{N}(k+1)X_i(k)}{\sum_{k=1}^{N} X_i(k)} \tag{2}$$

where $X_i(k)$, $k = 1, ..., N$ are the DFT coefficients of the i-th short-term frame and N is the number of frequency bins.

SC is directly related to the sharpness (high-frequency content) of the sound spectrum. Hence, higher SC values mean that there is a very bright sound with high frequencies present. The mean and median of the SC for each recording were calculated and normalized between 0 and 1.

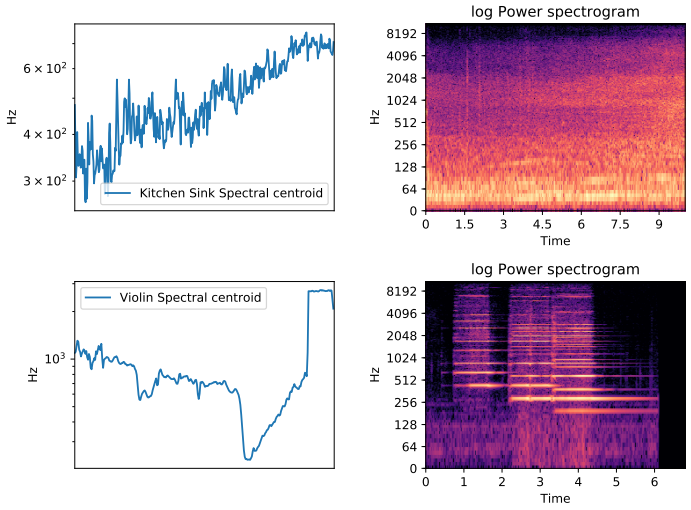


Figure 4: SC comparison between the sound of the kitchen sink (top) and the sound of a violin (bottom). The x-axis on shows time in s

Figure 4 shows a significant difference between the brighter sound of a violin and the more broadband sound of the running water of a kitchen sink. More specifically, for the kitchen sink, where low frequencies are mainly present, the

mean is 0.126 and the median is 0.113. On the other hand, the "sharper" sound of the violin, where the harmonics are very distinct at higher frequencies has a mean of 0.383 and a median of 0.366.

3.3.6. GFCC: Gammatone Frequency Cepstral Coefficients

The Gammatone filter-bank consists of a series of band-pass filters, which model the frequency selectivity property of the basilar membrane. The main difference between the MFCC and GFCC is that the Gammatone filter-bank and the cube root are used before applying the DCT while the triangular filter-bank and the log operation are applied in MFCC. Equation (3) describes the calculation of the GFCC:

$$GFCC_m = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} \log(E_n) \cos[\frac{\pi n}{N}(m - \frac{1}{2})], 1 \leq m \leq M \tag{3}$$

where E_n is the energy of the signal in the n-th band, N is the number of Gammatone filters and M is the number of GFCC.

We extracted 39 GFCC features per frame. These consisted of 13 GFCCs, 13 delta values and 13 delta-delta values.

3.4. Feature Selection

Feature selection was a crucial step for the first AED approach, since we wanted to have a framework that detects activities in real-time.

3.4.1. Feature Aggregation

Out of the 5,985 recordings (original=1,995 and two augmented=3,990), we extracted the following features: 173×16 (DWT) + 2 (ZCR) + 2 (SR) + 2 (SC) + 173×39 (GFCC) + 173×39 (MFCC). Aggregating all the features into a single vector is an important step before passing it to the sequential backward search algorithms and applying principal component analysis. Feature extraction and classification (using the first AED approach) ran on a Raspberry Pi 3 Model B platform.

3.4.2. SBS: Sequential Backward Selection

SBS starts from the whole feature set $X = \{x_i \mid i = 1, \ldots, N\}$ and discards the "worst" feature (x') at each step, such that the reduced set $X - \{x'\}$ gives the maximum value of an objective function $J(X - \{x'\})$. Given a feature set, SBS gives better results but is computationally more complex than other statistical feature selection methods [44]. With SBS, we reduced the number of features to 17 per recording.

3.4.3. PCA: Principal Component Analysis

The central idea of PCA is to reduce the dimensionality of a dataset that consists of many interrelated variables, while retaining as much as possible the variation present in the dataset. We applied PCA to the features given by SBS to reduce the feature space down to two principal components. The principal components were used as input to the classifier.

3.5. Activity Classification

For the first AED approach, we 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 a Gradient Boosting classifier.

For the second AED approach, we implemented a CNN based on a modified AlexNet [45] architecture. The CNN was trained on an NVIDIA GeForce GTX 1080 Ti.

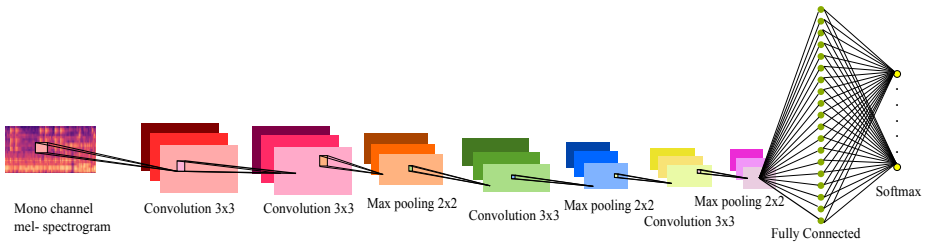


Figure 5: Second proposed AED approach

The CNN consists of 4 convolutional layers (Figure 5). The number of filters at each layer increases as a power of two. Specifically the first layer has 8 filters, the second 16, the third 32 and the fourth one 64. The first layer performs convolutions over the spectrogram of the input segment, using 3×3 kernels. The output is fed to a second convolutional layer which is identical to the first. A 2×2 max pooling operation follows the second layer and the subsampled feature maps are fed to two consecutive convolutional layers, each followed by max pooling operations. Each convolution operation is followed by batch normalization [46] of its outputs, before the element-wise application of the exponential linear unit (ELU) activation function [47] to facilitate training and improve convergence time. We selected the ELU activation function based on the results obtained by Clevert et al. [47], where it outperformed other commonly used activation functions (e.g., rectified linear unit (ReLU)), when tested on image datasets using deep neural networks with more than five layers. After each max pooling operation, we apply dropout [48] with an input dropout rate of 0.2. The number of kernels in all convolutional layers is 5. The resulting feature maps of the consecutive convolution-max pooling operations are then fed as input to a fully-connected layer with 128 logistic sigmoid units to which we also apply dropout with a rate of 0.2, followed by the output layer which computes the softmax function. Classification is obtained through hard assignment of the normalized output of the softmax function

$$c = \arg \max_{i=1,\ldots,N} y_i \tag{4}$$

$$y_i = \frac{\exp x_i}{\sum_{j=1}^{N} \exp x_j} \tag{5}$$

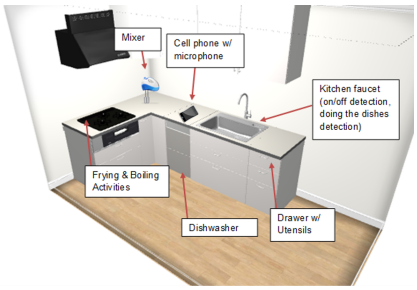
where N is the number of classes and x_i is the probability for the *i-th* class. We used the Adam optimizer [49] and trained our network with an initial learning rate l_r=0.001, which was reduced by a factor of 0.01 when there was no validation loss (categorical cross-entropy) improvement for five consecutive epochs. This ensured that there was no overfitting in the training. We trained the CNN

for 20 epochs.

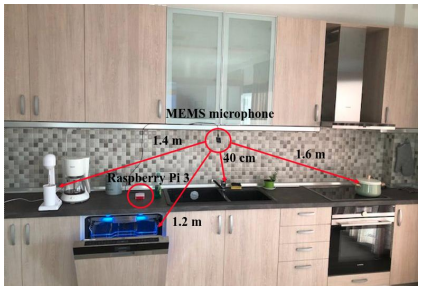
4. Experimental Setup

We recorded sounds of activities using the kitchen setup of Figure 6 (a), where there was no background noise and Figure 6 (b, c) that included background speech sounds and ambient noise of a fan and refrigerator. We also collected sounds for seven classes from Freesound [50].

(a) Home Kitchen Environment Setup



(b) CERTH Smart Home (KRIPIS) setup



(c) AKTIOS Kitchen Environment Setup

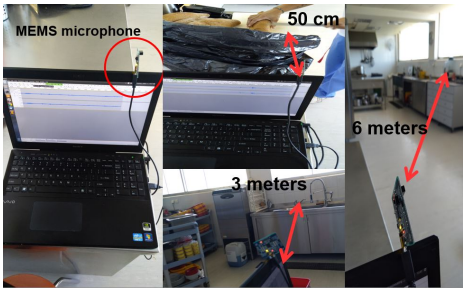


Figure 6: Experimental Setup

The first recordings were made in the kitchen of the first author (Figure 6 (a)). Only one person was present at the time of the recordings.For this environment, two smartphones (Samsung Galaxy S5 & ZTE Nubia Z11 miniS)

were placed on the kitchen counter above the dishwasher at an identical position. The main reason for using two smartphones was to capture the same source from two different, off the shelf, microphones. The smartphones were 50 cm away from the faucet, approximately 50 cm from the mixer, 1 m from the oven and approximately 2 m from the kitchen drawer. For the second set of recordings, the setup was as follows (Figure 6 (b)):

1. we used a Raspberry Pi 3 Model B with an MEMS DSP board to record the audio signals
2. the environment was noisier than for the first set of recordings because other researchers were present and speech or other environmental noises were captured more frequently
3. the MEMS board was placed at 1.2 m from the dishwasher, 1.4 m from the mixer, 40 cm from the kitchen faucet and 1.6 m from the oven. Compared to the first recordings, the distances to the appliances were larger to reduce the SNR

Finally, for the third set of recordings (Figure 6 (c)), we classified the audio signals in real-time using a laptop and an MEMS microphone board. For this environment there was a background noise of a fan and a refrigerator. We used the MEMS board to manually adjust the microphone gain (+6 dB; maximum threshold to avoid clipping when placed within 50 cm from the cutting board to detect the activity of bread cutting) for the recordings and the laptop to perform real-time classification. The MEMS board was placed in a fixed position on top of the laptop and 3 m from the kitchen faucet, 3 m from the dishwasher, 6 m from the mixer, and 50 cm from the cutting board.

A total of 1,995 audio signals from different activities were collected from the three kitchen environments (285 kitchen faucet, 285 boiling, 285 frying, 285 dishwasher, 285 mixer, 285 doing dishes and 285 cutting bread). All signals had a duration of 5 s. The setup included the following steps:

- we used data augmentation techniques as described in Section 3.2 to increase the total number of recordings in each class to 855

Table 1: Number of recordings of each class from different sources

Classes	**Kitchen Environment Figure 6(a)**	**Kitchen Environment Figure 6(b)**	**Kitchen Environment Figure 6(c)**	**Freesound**
Frying	160	85	-	40
Boiling	160	85	-	40
Mixer	160	40	45	40
Doing the dishes	160	85	-	40
Kitchen sink	160	34	51	40
Dishwasher	160	20	65	40
Cutting bread	-	-	285	-

- Monte Carlo cross-validation was used to randomly split the dataset into training and testing data (80% training and 20% testing) and the results (accuracy, precision, recall, F1-score) were averaged over the splits

The number of recordings for each class is summarized in Table 1.

5. Results

In this section, we present experiments to assess the performance of our two AED systems. In all experiments, we used 80% of the dataset for training and 20% for testing. For all classifiers, the split between training set and testing set was identical, as is common in the literature [51]. In Section 3.5.1, we compare several classifiers for the first system, we select the one with the highest F1-score and recognition accuracy and compare the performance to that of the second system. In Section 3.5.2, we study the effect of feature fusion on the recognition rate of the best classifier identified in Section 3.5.1 (Gradient Boosting). In Section 3.5.3, we study the recognition accuracy as a function of signal duration. In Section 3.5.4, we analyze the effect of both the SNR and distance between the microphone and event on the recognition accuracy in an "untrained" environment. In Section 3.5.5, we examine the response of the second AED system for an activity that was not included in the training set.

5.1. Selection of a traditional classifier for the first AED system and comparison with the second AED system

Table 2 compares the performance of various classifiers for the selected features. For all classifiers, we used the implementations in the scikit-learn [52] library. The signals used for this experiment were all the recordings from the three environments mentioned in Section 3.4 in addition to the Freesound recordings.

Table 2: Classifier Performance Comparison

MFCC+GFCC+SR+SC+ZCR+DWT (with augmented data)				
Classifier	**PRECISION**	**RECALL**	**F1-SCORE**	**ACCURACY**
kNN (5 nearest neighbors)	78.4%	79.4%	78.9%	79.4%
SVM (linear kernel)	79%	81.2%	80.1%	83.5%
SVM (RBF kernel)	84.1%	90.1%	87%	90.9%
Extra Trees	83.4%	85%	84.2%	89.7%
Random Forest	88.5%	89.1%	88.8%	91%
Gradient Boosting	***90.4%***	***90%***	***90.2%***	***91.7%***
Mel-Spectrogram (with augmented data)				
2D CNN /w 2D Max-pooling	***94.6%***	***90.9%***	***92.7%***	***96%***
1D CNN /w 1D Max-pooling	90%	89.7%	89.8%	91.3%

For the Random Forest classifier, we noticed, as the theory suggests, that increasing the number of trees can give a better and more stable performance; hence there is a small possibility of overfitting. The number of leaves in the tree had to be small, in order to capture noisy instances in the training dataset. Therefore, we selected 50 samples for each leaf node. For the RBF-based SVM classifier, the highest values for all evaluation measures were found for $\sigma = 1$ and $C = 0.1$. The parameter σ of the RBF kernel handles the non-linear classification and parameter C trades off correct classification of training examples against maximization of the decision functions margin. Finally, for Gradient Boosting we picked 500 estimators. We used the deviance (logistic regression) loss for classification with probabilistic outputs, since we had a multi-class problem. Another important parameter that affected the classification performance was the learning rate. We tried all values from 0.01 to 0.1 with a 0.01 step and selected 0.05, as it provided the best results. We kept the rest of the parameters

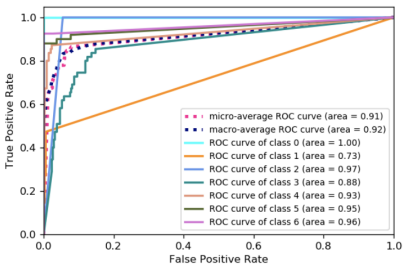
in the scikit-learn library at default settings. Additionally, as Gradient Boosting is fairly robust to overfitting, the large number of estimators resulted in a better performance, achieving an F1-Score of 90.2%. We obtained good results for boiling, frying, the use of the mixer, and also the use of the dishwasher. However, the activity of the "running" kitchen faucet was "understood" by our architecture as doing the dishes because some recordings were very similar due to the timing (meaning that no dishes or utensils were "heard" from the microphone).

We also applied McNemar's test to determine whether there was a significant difference between the accuracy of the classifiers. The results are summarized in Table 3 and show in particular that the 2D CNN classifier is statistically different from all other classifiers at the 0.05 significance level.

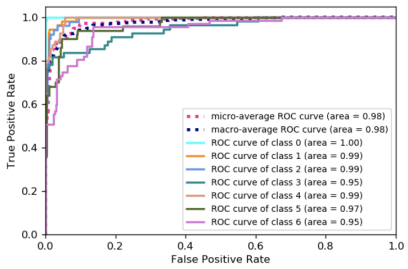
To further compare the performance of the classifiers, we plotted their Receiver Operating Characteristic (ROC) curves (Figure 7). We noticed that the boiling class was the most easily separable class for all the classifiers. The classes of cutting the bread and operating the kitchen faucet were the hardest ones for all the classifiers. This is because many recordings had sounds corresponding to these two particular classes towards the last second of the 5 s-recording.

In the following experiments, the first AED system was used with the Gradient Boosting classifier, since it achieved the highest performance characterized by a stable relationship between precision, recall, F1-Score and recognition accuracy.

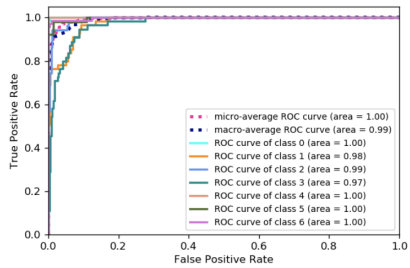
In order to highlight the importance of 2D max-pooling, we compared it to 1D max-pooling with a 1D CNN. The input to the 1D CNN network were mel-spectrograms with 128 bins. The resulting feature matrix input vector to the 1D CNN consisted of 128 mel-band energies in 431 successive frames (number of Fast Fourier Transform (FFT) samples = 1024 with hop length = 512, or window size of 20 ms with a 10 ms hop size from the next one). The 1D CNN had the same number of filters, kernels, etc. as the 2D CNN (described in Section 3.5). The main differences between the two networks are that kernels change from 3×3 to 3×1, max-pooling from 2×2 to 2×1 and in the Keras [53] library the $Conv2D$ and $MaxPooling2D$ layers are replaced with the $Conv1D$



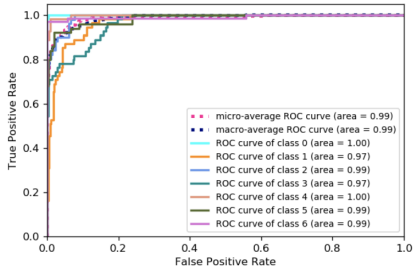
(a) kNN



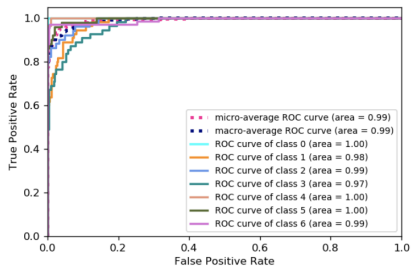
(b) SVM w/ linear kernel



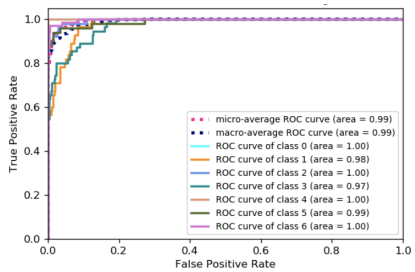
(c) SVM w/ RBF kernel



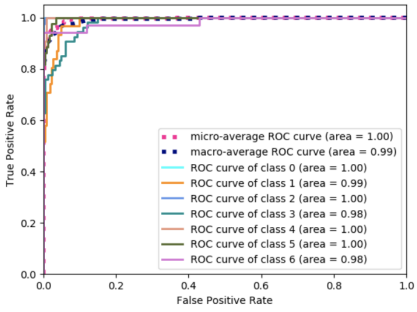
(d) Extra Trees



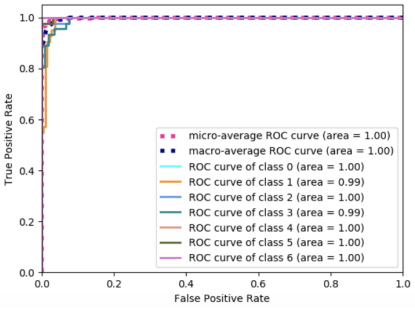
(e) Random Forest



(f) Gradient Boosting



(g) 1D CNN



(h) 2D CNN

Figure 7: ROC curves for the selected classifiers. Classes 0, 1, 2, 3, 4, 5 and 6 correspond to boiling, cutting bread, dishwasher, doing the dishes, frying, operating the kitchen faucet and mixer, respectively

and $MaxPooling1D$, respectively. The 2D CNN with 2D max-pooling, was able to capture the spatio-temporal information of the given signal and achieved an F1-Score of 92.7%. On the other hand, the 1D CNN achieved an F1-Score of 89.8% only. This showed that the audio signals that were present in the kitchen environment contained important information in the frequency domain.

5.2. Fusion of features for the first AED approach

Figure 8 shows how fusing features improves the performance of the first AED approach with the Gradient Boosting classifier. The accuracy rates were calculated for seven feature combinations.

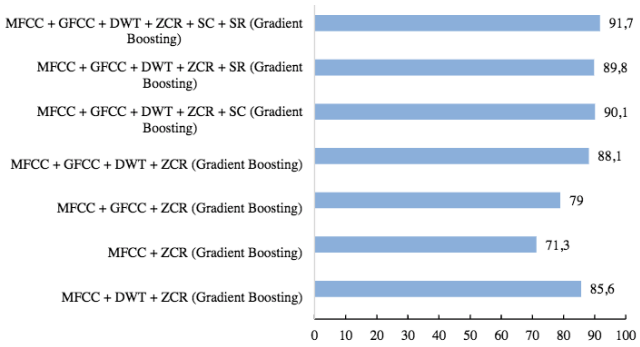


Figure 8: Recognition accuracy for different audio features using Gradient Boosting

Many sounds in a kitchen environment have an interchangeable pattern (bigger/smaller values for odd/even MFCCs). Some mechanical noises (mixer, dishwasher) have high short-time energy on their fundamental frequency and others (forks, spoons, trays) have high short-time energy on higher frequencies. This served as our motivation to test more time-frequency features in the kitchen recordings. Specifically, when introducing the GFCCs and the DWT, the recognition accuracy was significantly improved. MFCCs and ZCR achieved an accuracy of 71.3%. When we added the GFCCs first and DWT second, the accuracy improved to 79% and 85.6% respectively. GFCCs use the Equivalent Rectangular Bandwidth (ERB) scale. The ERB scale has a finer resolution at low frequencies, which were present in the kitchen environment, compared to the

mel-scale used by the MFCCs. Additionally, the DWT was able to separate the fine details of the input signal and increased the recognition accuracy. As for the classifiers, we applied McNemar's test on the features (Table 4) to check the statistical significance of the results.

5.3. Recognition accuracy as a function of the audio sample duration

We studied the impact of segment duration on the accuracy of activity recognition within the kitchen environment. Figure 9 shows that a 3 s time duration of the input signal is sufficient for accurate activity recognition. For the Gradient

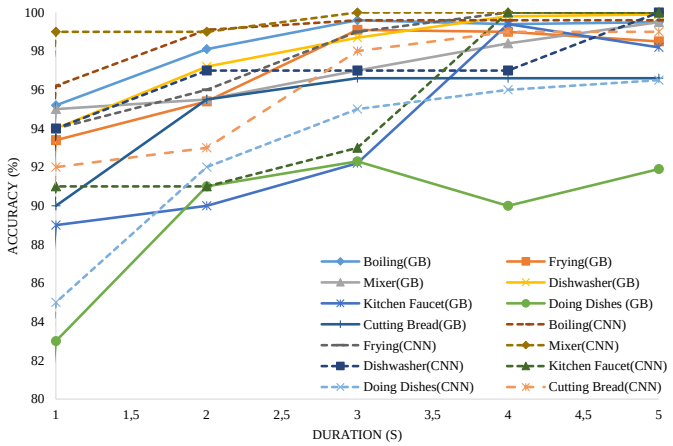


Figure 9: Recognition accuracy (using the Gradient Boosting classifier and the CNN) as a function of the sample duration

Boosting classifier, we noticed an unexpected drop-off for the activity of doing the dishes after three seconds. Examination of the confusion matrices revealed that there is a recognition uncertainty of 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 when the faucet was turned on and only at the last second of the recording an object (plate, utensils) was picked to be washed. On the other hand, the performance of the CNN improved as the audio clips became longer, since it was able to find clear patterns in the mel-spectrogram image.

5.4. Dependence of recognition accuracy on certain distance and SNR in a new environment

We trained both systems in the environments of Figure 6 (a-b) and tested them in the environment of Figure 6 (c). Our training set consisted of 1547 recordings for the seven classes. We tested the systems on the following classes only: *dishwasher, mixer, utensils/trays, kitchen faucet*. For this experiments, we renamed the class *doing the dishes* to *utensils/trays*, since the people in the kitchen rinsed the utensils/trays for a very short period of time and then used the dishwasher. 41 recordings for the activity of moving the utensils/trays were collected in order to test the two AED systems.

For this experiment we could not test all seven classes since, i) the recordings from cutting the bread were collected and trained using the setup of that environment, ii) there was no frying activity due to dietary instructions from the elderly care home where the experiment took place and iii) the setup was similar to a restaurant kitchen setup and we could not detect the boiling activity (the microphone was placed at a large distance from the stove). The results (Table 5 and Table 6) show that even with a relatively small training dataset and a distance of 3 m from the event to be classified, we were able to obtain satisfactory results when testing in a new indoor environment.

The distance between the activity and the microphone affected the recognition accuracy. Table 7 shows the SNR and the classification accuracy, using the Gradient Boosting and the CNN, of a set of activities at various distances. The ambient noise of the kitchen at AKTIOS (fan and refrigerator at -32 dB) operating at the time of the experiment dropped the performance of the approaches when increasing the distance from the microphone. The CNN outperformed the Gradient Boosting except when the mixer was used and the microphone was placed 6 m and 3 m away or when the dishwasher was used and the microphone was placed 1 m away.

5.5. Tests with activity that was not included in the training set (coffee machine) using the second AED system

For this experiment, we used the MEMS microphone board and a laptop 50 cm away from a coffee machine to collect 25 recordings of 5 s each. Out of the 25 recordings, 8 were classified as boiling, since it was the closest match in terms of the audio characteristics of the filter coffee machine. For the remaining 17, the classifier output was discarded because the output probability for each class was below the minimum threshold, set for this experiment to 0.7. More precisely, the class probability was between 0.5 and 0.6 for the boiling class and randomly distributed among the other classes.

6. Conclusions

We proposed two systems for AED in real-world conditions. The first one relies on feature extraction, selection, and classification, while the second one uses a CNN to learn from mel spectrogram images without the need for human-crafted features. Adding more audio features does not necessarily increase the recognition accuracy of the first system. However, feature selection methods and feature dimensionality reduction techniques, are critical to the success of the system. GFCCs and DWT coefficients significantly increased the recognition accuracy. They outperformed other well-known time-frequency features in the presence of background noise. Furthermore, we found that a signal duration of 3 s provided a good trade-off between time delay and recognition accuracy. The two systems were tested in a new environment and provided recognition accuracies above 90% for appliances that were up to 6 m away. This is a positive result since in most commercial kitchen environments, the distance between the microphone and the target appliance will be smaller.

Finally, in order to check the robustness of our second AED system, we tested it on an activity that was not included in the training set. The system correctly rejected the recording in 68% of the cases and misclassified it as boiling in the remaining cases.

The main limitation of the proposed systems is their inability to distinguish between overlapping events. Since only one acoustic sensor was used, only the loudest event was identified. For instance, when the microphone was placed 6 m away from the mixer and at the same time 3 m away from the kitchen faucet, the systems were able to correctly classify only the activity of the mixer, since the sound of the mixer masked entirely the sound of the running tap water. As future work, we will investigate to which extent multi-channel acoustic recordings are beneficial for the detection of domestic activities in different home environments. To this end, we will use the SINS database [54] which contains more than 200 hours of multi-channel recordings from different rooms (living room, kitchen, bathroom, bedroom). Additionally, we plan to keep collecting data in different rooms (e.g., living room, bathroom, etc.), introduce more effects such as reverberation and echo, and make the collected feature dataset publicly available, in order to help researchers working in this field evaluate their algorithms.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676157. The authors would also like to thank AKTIOS S.A. Elderly Care Units for allowing real-life testing of the framework.

References

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

[2] N.-C. Chi, G. Demiris, A systematic review of telehealth tools and interventions to support family caregivers, Journal of Telemedicine and Telecare 21 (1) (2015) 37–44.

[3] O. D. Lara, M. A. Labrador, A survey on human activity recognition using wearable sensors., IEEE Communications Surveys and Tutorials 15 (3) (2013) 1192–1209.

[4] Y. Lu, Y. Wei, L. Liu, J. Zhong, L. Sun, Y. Liu, Towards unsupervised physical activity recognition using smartphone accelerometers, Multimedia Tools and Applications 76 (8) (2017) 10701–10719.

[5] J. Yin, Q. Yang, J. J. Pan, Sensor-based abnormal human-activity detection, IEEE Transactions on Knowledge and Data Engineering 20 (8) (2008) 1082–1090.

[6] L. Meng, C. Miao, C. Leung, Towards online and personalized daily activity recognition, habit modeling, and anomaly detection for the solitary elderly through unobtrusive sensing, Multimedia Tools and Applications 76 (8) (2017) 10779–10799.

[7] A. Bulling, U. Blanke, B. Schiele, A tutorial on human activity recognition using body-worn inertial sensors, ACM Computing Surveys (CSUR) 46 (3) (2014) 33.

[8] Y. Chen, C. Shen, Performance analysis of smartphone-sensor behavior for human activity recognition, IEEE Access 5 (2017) 3095–3110.

[9] A. J. Perez, S. Zeadally, Privacy issues and solutions for consumer wearables, IT Professional 20 (4) (2018) 46–56.

[10] P. Kumari, L. Mathew, P. Syal, Increasing trend of wearables and multimodal interface for human activity monitoring: A review, Biosensors and Bioelectronics 90 (2017) 298–307.

[11] D. Giakoumis, G. Stavropoulos, D. Kikidis, M. Vasileiadis, K. Votis, D. Tzovaras, Recognizing daily activities in realistic environments through depth-based user tracking and hidden conditional random fields for mci/ad support, in: European Conference on Computer Vision, Springer, 2014, pp. 822–838.

[12] I. Kostavelis, D. Giakoumis, S. Malassiotis, D. Tzovaras, Human aware robot navigation in semantically annotated domestic environments, in: International Conference on Universal Access in Human-Computer Interaction, Springer, 2016, pp. 414–423.

[13] J. Chen, A. H. Kam, J. Zhang, N. Liu, L. Shue, Bathroom activity monitoring based on sound, in: International Conference on Pervasive Computing, Springer, 2005, pp. 47–61.

[14] M. Vacher, D. Istrate, F. Portet, T. Joubert, T. Chevalier, S. Smidtas, B. Meillon, B. Lecouteux, M. Sehili, P. Chahuara, et al., The sweet-home project: Audio technology in smart homes to improve well-being and reliance, in: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, IEEE, 2011, pp. 5291–5294.

[15] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, A. Sciarrone, Gender-driven emotion recognition through speech signals for ambient intelligence applications, IEEE Transactions on Emerging Topics in Computing 1 (2) (2013) 244–257.

[16] K.-Y. Huang, C.-C. Hsia, M.-s. Tsai, Y.-H. Chiu, G.-L. Yan, Activity recognition by detecting acoustic events for eldercare, in: 6th World Congress of Biomechanics (WCB 2010). August 1-6, 2010 Singapore, Springer, 2010, pp. 1522–1525.

[17] M. Sehili, B. Lecouteux, M. Vacher, F. Portet, D. Istrate, B. Dorizzi, J. Boudy, Sound environment analysis in smart home, Ambient Intelligence (2012) 208–223.

[18] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, M. Omologo, Acoustic event detection and classification, in: Computers in the human interaction loop, Springer, 2009, pp. 61–73.

[19] H. Lozano, I. Hernáez, A. Picón, J. Camarena, E. Navas, Audio classification techniques in home environments for elderly/dependant people, in:

International Conference on Computers for Handicapped Persons, Springer, 2010, pp. 320–323.

[20] J. Chen, A. H. Kam, J. Zhang, N. Liu, L. Shue, Bathroom activity monitoring based on sound, in: Proceedings of the Third International Conference on Pervasive Computing, PERVASIVE'05, Springer-Verlag, Berlin, Heidelberg, 2005, pp. 47–61.

[21] F. Kraft, R. Malkin, T. Schaaf, A. Waibel, Temporal ICA for classification of acoustic events in a kitchen environment, in: INTERSPEECH, Lisbon, Portugal, 2005.

[22] R. M. Alsina-Pagès, J. Navarro, F. Alías, M. Hervás, homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring, Sensors 17 (4) (2017) 854.

[23] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, N. Bonnefond, The sweet-home speech and multimodal corpus for home automation interaction, in: The 9th edition of the Language Resources and Evaluation Conference (LREC), 2014, pp. 4499–4506.

[24] S. Chu, S. Narayanan, C.-C. J. Kuo, Environmental sound recognition with time–frequency audio features, IEEE Transactions on Audio, Speech, and Language Processing 17 (6) (2009) 1142–1158.

[25] D. Barchiesi, D. Giannoulis, D. Stowell, M. D. Plumbley, Acoustic scene classification: Classifying environments from the sounds they produce, IEEE Signal Processing Magazine 32 (3) (2015) 16–34.

[26] G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, et al., Convolutional recurrent neural networks for polyphonic sound event detection, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (6) (2017) 1291–1303.

[27] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, G. Widmer, CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep

convolutional neural networks, IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE).

[28] J. Liu, X. Yu, W. Wan, C. Li, Multi-classification of audio signal based on modified svm, in: IET International Communication Conference on Wireless Mobile and Computing (CCWMC 2009), 2009, pp. 331–334.

[29] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. B. Jackson, M. D. Plumbley, Unsupervised feature learning based on deep models for environmental audio tagging, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (6) (2017) 1230–1241.

[30] J. Li, W. Dai, F. Metze, S. Qu, S. Das, A comparison of deep learning methods for environmental sound detection, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 126–130.

[31] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, A. Serralheiro, Non-speech audio event detection, in: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, IEEE, 2009, pp. 1973–1976.

[32] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, J. Huopaniemi, Audio-based context recognition, IEEE Transactions on Audio, Speech, and Language Processing 14 (1) (2006) 321–329.

[33] J. T. Geiger, B. Schuller, G. Rigoll, Large-scale audio feature extraction and svm for acoustic scene classification, in: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on, IEEE, 2013, pp. 1–4.

[34] F. Fuhrmann, A. Maly, C. Leitner, F. Graf, Three experiments on the application of automatic speech recognition in industrial environments, in:

International Conference on Statistical Language and Speech Processing, Springer, 2017, pp. 109–118.

[35] H. Erdogan, J. R. Hershey, S. Watanabe, J. Le Roux, Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio, in: New Era for Robust Speech Recognition, Springer, 2017, pp. 165–186.

[36] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, W. Xiao, H. Phan, Continuous robust sound event classification using time-frequency features and deep learning, PloS one 12 (9).

[37] C.-Y. Wang, J.-C. Wang, A. Santoso, C.-C. Chiang, C.-H. Wu, Sound event recognition using auditory-receptive-field binary pattern and hierarchical-diving deep belief network, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (8) (2018) 1336–1351.

[38] J. Salamon, J. P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification, IEEE Signal Processing Letters 24 (3) (2017) 279–283.

[39] B. McFee, E. J. Humphrey, J. P. Bello, A software framework for musical data augmentation., in: ISMIR, 2015, pp. 248–254.

[40] B. Milner, J. Darch, I. Almajai, 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, 2008, pp. 1–5.

[41] L. Rabiner, Fundamentals of speech recognition, PTR Prentice Hall, 1993.

[42] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, in: Proceedings of the 14th python in science conference, 2015, pp. 18–25.

[43] S. Bilgin, O. Polat, 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, 2009, pp. 1–4.

[44] A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2) (1997) 153–158.

[45] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[46] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.

[47] D.-A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (eLUs), arXiv preprint arXiv:1511.07289.

[48] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting., Journal of machine learning research 15 (1) (2014) 1929–1958.

[49] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceedings of the 3rd International Conference for Learning Representations (ICLR-15), 2014.

[50] Robinhood76, Kitchen common sounds (2008).
URL `https://www.freesound.org/people/Robinhood76/packs/3870`

[51] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Transactions on Medical Imaging 35 (5) (2016) 1285–1298.

[52] scikit-learn. machine learning in python (2019).
URL https://scikit-learn.org/stable

[53] F. Chollet, et al., Keras, https://github.com/fchollet/keras (2015).

[54] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, P. Karsmakers, The SINS database for detection of daily activities in a home environment using an acoustic sensor network, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), 2017, pp. 32–36.

Table 3: McNemar's test results

Classifiers	p-value	Statistically Significant ($p < 0.05$)
kNN vs SVM Linear	0.01337	Yes
kNN vs SVM RBF	<0.001	Yes
kNN vs Extra Trees	<0.001	Yes
kNN vs Random Forest	<0.001	Yes
kNN vs Gradient Boosting	<0.001	Yes
kNN vs 2D CNN	<0.001	Yes
kNN vs 1D CNN	<0.001	Yes
SVM linear vs SVM RBF	<0.001	Yes
SVM linear vs Extra Trees	<0.001	Yes
SVM linear vs Random Forest	<0.001	Yes
SVM linear vs Gradient Boosting	<0.001	Yes
SVM linear vs 2D CNN	<0.001	Yes
SVM linear vs 1D CNN	<0.001	Yes
SVM RBF vs Extra Trees	0.52239	No
SVM RBF vs Random Forest	0.86793	No
SVM RBF vs Gradient Boosting	0.51137	No
SVM RBF vs 2D CNN	<0.001	Yes
SVM RBF vs 1D CNN	0.74282	No
Extra Trees vs Random Forest	0.24778	No
Extra Trees vs Gradient Boosting	0.05247	No
Extra Trees vs 2D CNN	<0.001	Yes
Extra Trees vs 1D CNN	0.21532	No
Random Forest vs Gradient Boosting	0.62905	No
Random Forest vs 2D CNN	<0.001	Yes
Random Forest vs 1D CNN	1	No
Gradient Boosting vs 2D CNN	0.00259	Yes
Gradient Boosting vs 1D CNN	0.82380	No
2D CNN vs 1D CNN	<0.001	Yes

Table 4: McNemar's test on the features

Features	**p-value**	**Statistically Significant (p <0.05)**
MFCCs + GFCCs + DWT + ZCR + SC + SR vs MFCCs + GFCCs + DWT + ZCR + SR	0.23788	No
MFCCs + GFCCs + DWT + ZCR + SC + SR vs MFCCs + GFCCs + DWT + ZCR + SC	0.14346	No
MFCCs + GFCCs + DWT + ZCR + SC + SR vs MFCCs + GFCCs + DWT + ZCR	0.01612	Yes
MFCCs + GFCCs + DWT + ZCR + SC + SR vs MFCCs + GFCCs + ZCR	<0.001	Yes
MFCCs + GFCCs + DWT + ZCR + SC + SR vs MFCCs + ZCR	<0.001	Yes
MFCCs + GFCCs + DWT + ZCR + SC + SR vs MFCCs + DWT + ZCR	<0.001	Yes
MFCCs + GFCCs + DWT + ZCR + SR vs MFCCs + GFCCs + DWT + ZCR + SC	1	No
MFCCs + GFCCs + DWT + ZCR + SR vs MFCCs + GFCCs + DWT + ZCR	0.24778	No
MFCCs + GFCCs + DWT + ZCR + SR vs MFCCs + GFCCs + ZCR	<0.001	Yes
MFCCs + GFCCs + DWT + ZCR + SR vs MFCCs + ZCR	<0.001	Yes
MFCCs + GFCCs + DWT + ZCR + SR vs MFCCs + DWT + ZCR	0.01609	Yes
MFCCs + GFCCs + DWT + ZCR + SC vs MFCCs + GFCCs + DWT + ZCR	0.16863	No
MFCCs + GFCCs + DWT + ZCR + SC vs MFCCs + GFCCs + ZCR	<0.001	Yes
MFCCs + GFCCs + DWT + ZCR + SC vs MFCCs + ZCR	<0.001	Yes
MFCCs + GFCCs + DWT + ZCR + SC vs MFCCs + DWT + ZCR	0.00642	Yes
MFCCs + GFCCs + DWT + ZCR vs MFCCs + GFCCs + ZCR	<0.001	Yes
MFCCs + GFCCs + DWT + ZCR vs MFCCs + ZCR	<0.001	Yes
MFCCs + GFCCs + DWT + ZCR vs MFCCs + DWT + ZCR	0.15385	No
MFCCs + GFCCs + ZCR vs MFCCs + ZCR	0.00239	Yes
MFCCs + GFCCs+ ZCR vs MFCCs + DWT + ZCR	0.00254	Yes
MFCCs + ZCR vs MFCCs + DWT + ZCR	<0.001	Yes

Table 5: Confusion matrix using Gradient Boosting for the classes of the framework in a new environment (not included in the training dataset). The distance between the microphone and each activity was 3 m

	Mixer	Dishwasher	Utensils/Trays	Kitchen Faucet
Mixer	45	11	0	1
Dishwasher	0	48	2	10
Utensils/Trays	0	1	39	0
Kitchen Faucet	0	5	0	40

Table 6: Confusion matrix using CNN for the classes of the framework in a new environment (not included in the training dataset). The distance between the microphone and each activity was 3 m

	Mixer	Dishwasher	Utensils/Trays	Kitchen Faucet
Mixer	45	11	0	0
Dishwasher	0	46	0	5
Utensils/Trays	0	2	41	0
Kitchen Faucet	0	6	0	46

Table 7: Recognition accuracy of Gradient Boosting and CNN according to distances and SNRs

Activities	**Distance (m)**	**SNR (dB)**	**Accuracy with Gradient Boosting (%)**	**Accuracy with CNN (%)**
Using the Kitchen Sink	3	-27	90.2	*93.4*
	0.4	-10	94	*98.8*
Using the Mixer	6	-11	*98.5*	97.1
	3	-8	100	100
Moving the Utensils/Trays	6	-16	91.1	*95*
	3	-13	96.8	*100*
Using the Dishwasher	3	-30	*90.2*	89.9
	1	-25	*93*	91.7