ISSN 2007-9737

A Comparative Study in Machine Learning and Audio Features for Kitchen Sounds Recognition

Alain Manzo-Martínez, Fernando Gaxiola, Graciela Ramírez-Alonso, Fernando Martínez-Reyes

Universidad Autónoma de Chihuahua,
Facultad de Ingeniería,
Mexico

{amanzo, lgaxiola, galonso, fmartine}@uach.mx

Abstract. For the last decades the work on audio recognition has been directed to speech and music, however, an increasing interest for the classification and recognition of acoustic events is observed for the last years. This poses the challenge to determine the identity of sounds, their sources, and the importance of analysing the context of the scenario where they act. The aim of this paper is focused on evaluating the robustness to retain the characteristic information of an acoustic event against the background noise using audio features in the task of identifying acoustic events from a mixture of sounds that are produced in a kitchen environment. A new database of kitchen sounds was built by us, since in the reviewed literature there is no similar benchmark that allows us to evaluate this issue in conditions of 3 decibels for the signal to noise ratio. In our study, we compared two methods of audio features, Multiband Spectral Entropy Signature (MSES) and Mel Frequency Cepstral Coefficients (MFCC). To evaluate the performance of both MSES and MFCC, we used different classifiers such as Similarity Distance, k-Nearest Neighbors, Support Vector Machines and Artificial Neural Networks (ANN). The results showed that MSES supported with an ANN outperforms any other combination of classifiers with MSES or MFCC for getting a better score.

Keywords. Entropy, neural networks, mixture of sounds, MFCC.

1 Introduction

Sounds are around human being everywhere and due to physic properties of the sounds one can heard the acoustics of these. Acoustic events refer to several everyday sounds which are generated in natural or artificial form (namely, the sounds found in the environment of the everyday life, excluding speech and music). The development of an acoustic event recognition system (AERS), contributes to the development of intelligent systems capable to understand sound within a context. These systems are important for real-world applications such as activity monitoring systems [15, 39, 45], ambient assisted living [28, 40, 50], human-computer interaction [8, 41, 42], security surveillance [1, 2], assisted robotics [21, 38, 48], among others.

Automatic recognition of acoustic events in real situations is not an easy task, because the audio captured by microphones contains a mixture of different sources of sound. Recent research work about AERS has focused on two types of classification problems: acoustic events classification for a specific context and acoustic events recognition into contextual classes [52, 53]. The former can be associated to, for instance, the activity recognition in a home environment, where acoustic events can offer information that occur in a specific dwelling space.

The audio information for scene understanding, can be more assertive if they exclusively recognize the acoustic events that occur in a specific place. On the other hand, there is the recognition of human activity from the sounds that occur in different places, for instance, the difference of contextual events between the home and the office. It would be difficult to say what kind of activity is carried out, if the contextual classes of the sounds

are not clear. Besides, not limiting the sounds in the scene will be even more difficult this task.

Preliminary work with AERS adopted approaches used for the processing of speech and music, however, the non-stationary characteristics of the acoustic events made the recognition of events problematic for databases with a great number of sound sources [13]. For example, in speech recognition is common to use a phonetic structure that can be seen as a basic component of voice, therefore, spoken words can be divided into elemental phonemes over which it is possible the application of probabilistic models. Conversely, phoneme based approaches cannot be applied to acoustic events coming from sounds created by a car crash or due to pouring water in a glass. Even, if it is possible to create a dictionary of basic units of these events, modelling signal variation in time would be difficult. The same occurs whit the attempt to compare music and acoustic events because the latter does not exhibit significant stationary patterns such as melody and rhythm [13].

The recognition of acoustic events involves two phases: a feature extraction phase followed by a classification phase. The feature extraction phase allows to play two roles; a dimension reduction role, and a representation role. An AERS uses stationary and non-stationary feature extraction techniques. Most of the features extraction algorithms use a scheme called bag-of-frames. The bag-of-frame approach consists of considering the signal in a blind way, using a systematic and general scheme where the signal is divided into consecutive overlapping frames, from which a vector of features is determined. The features are supposed to represent characteristic information of the signal for the problem at hand. These vectors are then aggregated (hence the “bag”) and fed to the next phase of an audio recognition system [3].

Audio signals have been traditionally characterized by Mel Frequency Cepstral Coefficients (MFCC). The methodology for computing MFCC involves a filter bank that approximates some important properties of the human auditory system. MFCC has been shown to work well for structured sounds such as speech and music [16, 23, 25, 26, 27, 37]. Since MFCC has been successfully used in speech and music applications, some work suggests the use of MFCC for characterizing acoustic events that contains a large and diverse variety of sounds, including those with strong temporal domain [4, 35, 40]. In addition, MFCC are often used by researchers for benchmarking their works.

For the classification phase of an AERS there are different machine learning techniques such as Support Vector Machine (SVM). SVM is a classifier that discriminates the data by creating boundaries between classes rather than estimating class conditional densities, or in other words, that SVM could draw accurate classification rate even if the sample size is small, a common scenario for acoustic event classification [14, 24, 51].

Artificial Neural Networks (ANN) is another machine learning technique being widely used for audio recognition systems. ANN deals with the study and construction of systems able to learn from the data. ANN algorithms infer unknowns from known data a characteristic that might describe acoustic events where there is an acoustic event of interest that need to be differentiated from a mix of sounds. [7, 29, 36].

There are other techniques that can be used to identify acoustic events such as audio signatures recognition. In these technique the challenge is to find the acoustic events that sound similar to the audio that the system captures. The similarity rate is evaluated using a distance function. Audio signatures use two fundamental processes to be determined, a feature extraction process and a modeling process. The latter refers to the minimal compacted representation that can be achieved to describe a signal, but which robustness preserve the model against typical audio degradation [40]. Audio signatures thus work very well on AERS, but the problem is complicated when it is required to identify an acoustic event present in a mixture of sounds. This problem usually leads to apply source separation techniques and machine learning algorithms to treat with the complexity of the signals.

In this work we considered the signals unprocessed. Also, we use no source separation technique because our intention is to evaluate the robustness to retain the characteristic information

of an acoustic event against the background noise using two audio features, MFCC that is the state–of–the–art benchmark and the multiband spectral entropy signature (MSES), a technique that has been successfully used in audio fingerprinting, speech recognition and others applications of audio [6, 9, 10, 11, 12, 30].

In addition, MSES feature has never been studied to recognize acoustic events exclusively for indoor domestic environments. For the previously mentioned, the audio signature approach is used, namely, it is assumed that only there is one instance per acoustic event (for the traditional audio signature approach, only there is one version of the songs) for the types of sound classes to be considered and versions contaminated with noise of that instance (it is similar to distort each song with different types of degradations). Therefore, the aim is not to classify different instances of acoustic events into classes, but to evaluate robustness of MFCC versus MSES using a low level of SNR (Signal to Noise Ratio) in the mix of acoustic events.

The machine learning techniques utilized in this paper were selected according to the results in recognition and classification issues of related literature, besides the configuration of them are performed follow the experimental ideas in that literature and in some cases with optimization algorithms [7, 14, 21, 26].

Regarding classification, an optimization with genetic algorithm and particle swarm optimization were developed in order to improve the performance of the best combination achieved between audio features and the studied classifiers.

The built database is an additional contribution since there is no database in the reviewed works similar to the one that we propose in this paper. It has the particularity of being complex in its construction by mixing sounds at a low level of SNR. Forward, we describe in detail this database and we encourage to the readers to use it in their future works. In our case, it will be part of our tested towards exploring recognition of activity for elders living alone, for instance, to identify acoustic of events that might indicate whether the elder is using the blender or for identifying sounds of risk in a home environment.

2 Theoretical Background

The characterization of audio signals is related to the process of extracting the characteristics that abstractly describe a signal and reflect their most relevant aspects of perception. To extract the characteristics of an audio signal, it is common to segment the signal in short frames, possibly overlapping it sufficiently close to each other, in such a way that multiple events distinguishable or perceptual are not covered in a single frame [3]. This process of splitting the signal into frames is a characteristic part for computing MFCC and MSES. The next subsections describe the process for determining both audio features, as well as the different classification techniques used in the experiments that support our results.

2.1 Mel Frequency Cepstral Coefficients

MFCC are short-term spectral-based features and its success have been due to their ability to represent the amplitude spectrum in a compact form. MFCC is based on the non-linear frequency scale of human auditory perception which use two types of filters, linearly spaced filters and logarithmically spaced filters. The signal is expressed in Mel's frequency scale to capture the most important characteristics of an audio [46].

For computing MFCC, the audio signal is divided into short time frames for extracting from each one a feature vector with L coefficients. We compute the Short Time Fourier Transform for each frame, which it is given by (1), for $k = 0, 1, ..., N-1$, where k correspond to the frequency $f(k) = kf_s/N$, and f_s is the sampling frequency in Hertz. Here, $x(n)$ denotes a frame of length N and $w(n)$ is the Hann window function which it is given by $w(n) = 0.5 + 0.5cos(2\pi n/N)$:

$$X(k) = \sum_{n=0}^{N-1} x(n)w(n)e^{-i2\pi kn/N}. \qquad (1)$$

The process continues scaling the magnitude spectrum $|X(k)|$ in both frequency and magnitude. First, the frequency is scaled using the so-called

Mel filter Bank $H(k,m)$ and then the logarithm is taken using (2):

$$X^{'}(m) = \ln\left(\sum_{k=0}^{N-1}|X(k)|H(k,m)\right), \qquad (2)$$

for $m = 1, 2, ..., M$, where M is the number of filters and $M \ll N$. The Mel filter bank, $H(k,m)$, is a set of triangular filters, where the frequencies in Mel scale of the filter bank are computed with $\phi = 2595log_{10}(f/700 + 1)$, which is a common approximation. MFCC are obtained decorrelating the spectrum $X^{'}(m)$ by computing the Discrete Cosine Transform using (3):

$$c(l) = \sum_{m=1}^{M} X^{'}(m)cos\left[l\frac{\pi}{M}\left(m - \frac{1}{2}\right)\right], \qquad (3)$$

for $l = 1, 2, ..., L$, where $c(l)$ is the lth MFCC. With this procedure, a vector with L coefficients is extracted from each frame.

In this work, we will focus on the ISP implementation for computing MFCC [46], this implementation considers a filter bank $H(k,m)$ with logarithmic spacing and constant amplitude, where the number of filters is a custom parameter.

2.2 Shannon's Entropy and Spectral Entropy

When the audio signals are severely degraded, the features that describe it usually disappear, therefore, the problem becomes finding the features that would still be present in the signal despite the level of degradation to which it was subjected. Authors focused on this problem have explored entropy to characterize audio signals as robustly as possible to different types of degradations. In this address, we will start by discussing about the Shannon's entropy and spectral entropy concept.

In information theory, Shannon's entropy is related to the uncertainty of a source of information [43]. For example, entropy is used to measure the predictability of a random signal and the "peakiness" of a probability distribution function. In research, it is common to use (4) to measure, through entropy, the amount of information the signal carries. Here, p_i is the probability for any sample of the signal to have value i being n the number of possible values the samples may adopt:

$$H = -\sum_{i=1}^{n} p_i ln(p_i). \qquad (4)$$

Some estimate of the Probability Distribution Function (PDF) is needed to determine the entropy of a signal, therefore, it can be used both parametric and non-parametric methods, and histograms. If histograms are chosen, we have to be careful that the amount of data involved is high enough to avoid peaks in the histogram.

When talking about spectral entropy it is necessary to review Shen's work [44], since that concept was introduced for the first time as an additional feature for endpoint detection (voice activity detection). The idea of spectral entropy compromises to consider the spectrum of a signal as a PDF to capture the peaks of the spectrum and their location. In order to convert the spectrum into a PDF, the individual frequency components of the spectrum are separated and divided by sum of all the components, namely, $p_k = X(k)/\sum_{i=1}^{N} X(i)$, for $k = 1, 2, ..., N$, where $X(k)$ is the energy of kth frequency component of the spectrum, $\mathbf{p} = (p_1, ..., p_N)$ is the PDF of the spectrum and N is the total number of frequency components in the spectrum. This ensures the PDF area is one and can be used for computing entropy.

The concept of multiband spectral entropy was introduced by [32], and it consists of dividing the spectrum into equal-sized sub-bands to compute entropy on each one of them by using (4), where each sub-band spectrum should be assumed a PDF. Additionally, [33] proved that the multiband spectral entropy works very well with additive wide-band noise and at low levels of SNR.

2.3 Multiband Spectral Entropy Signature

Based on the idea presented by Misra et al. [32, 33], spectral entropy concept can be used for getting a robust signature that can be useful in different audio recognition issues [6, 9, 10, 11, 12, 30]. Unlike Misra et al., this work compute entropy

at each sub-band by using the entropy of a random process [9].

Let $\mathbf{x} = [x_1, x_2, ..., x_n]^T$ be a vector of n real-valued random variables, then, $\mathbf{x}$ is said to be a Gaussian random vector where the random variables x_i are said to be jointly Gaussian if the joint probability density function of the n random variables x_i is given by $p(\mathbf{x}) = \mathcal{N}(\mathbf{m_x}, \mathbf{\Sigma_x})$, where $\mathbf{m_x} = [m_1, m_2, ..., m_n]^T$ is a vector containing the means of x_i, this is, $m_i = E[x_i]$. $\mathbf{\Sigma_x}$ is a symmetric positive definite matrix with elements σ_{ij} that are the covariances between x_i and x_j, this is, $\sigma_{ij} = E[(x_i - m_i)(x_j - m_j)]$.

Taking some precautions, the entropy of a Gaussian random vector can be determined using the continuous version of the Shannon's entropy, which is given by (5):

$$H(\mathbf{x}) = -\int_{-\infty}^{+\infty} p(\mathbf{x}) ln[p(\mathbf{x})] d\mathbf{x}. \tag{5}$$

If it is assumed that the random vector follows a Gaussian distribution with a mean of zero and the covariance matrix given by $N(0, \mathbf{\Sigma_x})$, then replacing $p(\mathbf{x})$ into (5), we get the equation for determining the entropy of a vector on a random process [34], equation (6), where $|\mathbf{\Sigma_x}|$ is the determinant of the covariance matrix:

$$H(\mathbf{x}) = \frac{n}{2} ln(2\pi e) + \frac{1}{2} ln(|\mathbf{\Sigma_x}|). \tag{6}$$

In order to compute MSES, the audio signal should be divided into frames, and for each of these to extract a vector with L coefficients of entropy. Next, the Short Time Fourier Transform is computed on each frame by using (1). For the division of the full-band spectrum into sub-bands, we take into account the idea about how people identify sounds. The human ear perceives better lower frequencies than higher ones, but not all frequencies can be heard with the same sensitivity. This process can be modeled in the whole bandwidth of the response of the ear using the Bark scale, which it is divided in 25 critical bands [49, 47]. Table 1 shows the first 24 critical bands with their respective bandwidths.

Table 1. Critical bands for the Bark scale

Critical Band	Lower cut-off (Hz)	Central Frequency (Hz)	Higher cut-off (Hz)	Bandwidth (Hz)
1	0	50	100	100
2	100	150	200	100
3	200	250	300	100
4	300	350	400	100
5	400	450	510	110
6	510	570	630	120
7	630	700	770	140
8	770	840	920	150
9	920	1000	1080	160
10	1080	1170	1270	190
11	1270	1370	1480	210
12	1480	1600	1720	240
13	1720	1850	2000	280
14	2000	2150	2320	320
15	2320	2500	2700	380
16	2700	2900	3150	450
17	3150	3400	3700	550
18	3700	4000	4400	700
19	4400	4800	5300	900
20	5300	5800	6400	1100
21	6400	7000	7700	1300
22	7700	8500	9500	1800
23	9500	10500	12000	2500
24	12000	13500	15500	3500

We use (7) to change Hertz to Barks, where f is the frequency in Hertz:

$$Barks = 13tan^{-1}\left(\frac{0.75f}{1000}\right) + 3.5tan^{-1}\left[\left(\frac{f}{7500}\right)^2\right]. \tag{7}$$

The process continues computing entropy for each one of the critical bands using (6). It was considered for each sub-band that spectral coefficients are distributed normally. This consideration is due to that a good estimate of the PDF cannot be determined by using non-parametric methods, since the lowest bands of the spectrum have too few coefficients. For computing entropy, a random process with two random variables was considered. Real and imaginary parts of the spectral coefficients are assumed to be random variables with a normal distribution and zero mean, hence, for the two-dimensional case the entropy is determined by $H = ln(2\pi e) + (1/2)ln(\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2)$, where σ_{xx} and σ_{yy} are the variances of the real and imaginary parts, respectively, and σ_{xy} is the covariance between the real and imaginary parts. The result of this process is a $L \times T$ matrix (named as signature), where L is the number of coefficients of entropy and T denotes the number of frames.

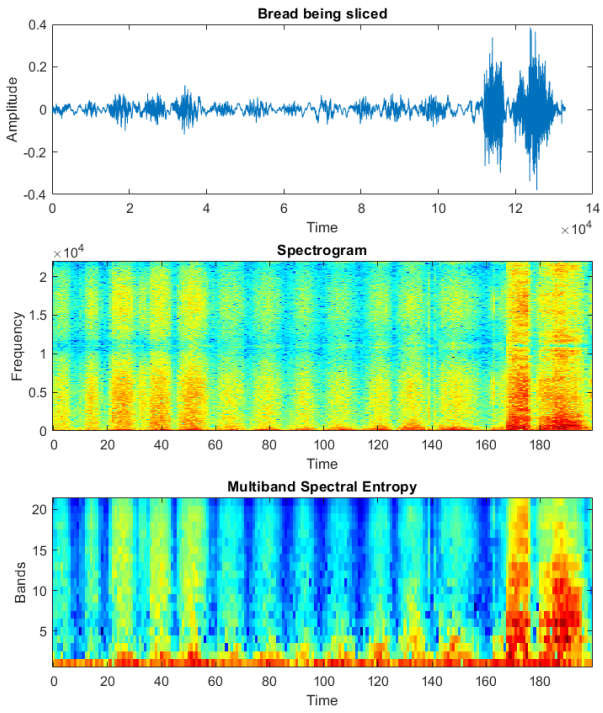

Fig. 1. Illustration of MSES signature with its corresponding signal and spectrogram. This signature corresponds to three seconds of audio from the acoustic event called "Bread being sliced"

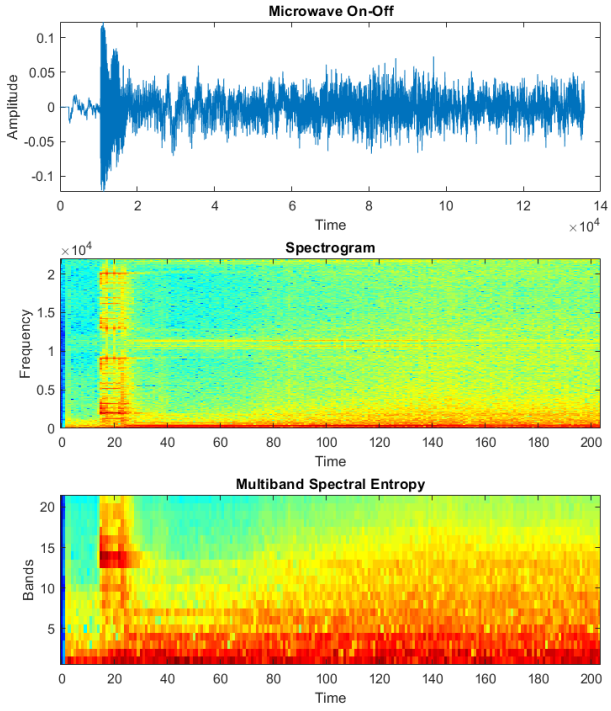

Fig. 2. Illustration of MSES signature with its corresponding signal and spectrogram. This signature corresponds to three seconds of audio from the acoustic event called "Microwave On-Off"

This signature captures the level of information content for every critical band and frame position in time.

Figures 1 and 2 show the signatures of two acoustic events that are obtained with the MSES method. The signals in time domain of the acoustic events "Bread being sliced" (Fig.1) and "Microwave On-Off" (Fig.2) are showed in the upper panels, whereas the spectrograms of both signals appears in the middle panels. The bottom of each one of the Figures displays the signatures for both acoustic events using MSES method.

2.4 Similarity Distance Functions

A measure of similarity indicates the strength of the relationship between two data points. The more the two data points resemble one another, the larger the similarity measure is. Let $\mathbf{x} = (x_1, x_2, ..., x_d)$ and $\mathbf{y} = (y_1, y_2, ..., y_d)$ be two d-dimensional data points. Then the similarity between $\mathbf{x}$ and $\mathbf{y}$ will be some function of their attribute values, as shown in (8):

$$s(\mathbf{x}, \mathbf{y}) = s(x_1, x_2, ..., x_d, y_1, y_2, ..., y_d). \quad (8)$$

A similarity distance function refers to a function $s(x, y)$ measured on any two arbitrary data points in a data set that satisfies the following properties [17]:

1. $0 \leq s(\mathbf{x}, \mathbf{y}) \leq 1$,
2. $s(\mathbf{x}, \mathbf{x}) = 1$,
3. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$.

The idea of similarity is more consistent if one considers the function of Hamming distance, since it determines the distance between two arbitrary

data points as the number of symbols or bits in which they differ. Another distance that is adopted to measure the similarity between two data points is the Cosine distance [17]. Cosine distance measures the similarity between two vectors in a space that has an internal product with which the value of the cosine of the angle between them is evaluated.

2.5 Artificial Neural Networks

The Artificial Neural Network (ANN) is a mathematical model that simulate the behavior of a biological neuron of humans. The ANN emulate the process of learning of the humans based at the equation (9). The approach for succeeding learning depends of the inputs x_i which they are multiplied with weights w_i, later, a transfer function $f(*)$ is applied for obtain the final result y for the ANN:

$$y = f\left(\sum_{i=1}^{n} x_i w_i\right). \tag{9}$$

The transfer function used in a neural network can be the sigmoidal function, linear function and hyperbolic tangent sigmoid function. For training the neural network is utilized the back-propagation method for update the weights in each epoch. The algorithm for learning can be the descendent gradient with the variants of learning rate, momentum and the use of both, also the scaled conjugate gradient and the variants of Fletcher-Reeves, Polak-Ribiére and Powell-Beale for the conjugate gradient.

2.6 Support Vector Machine

The Support Vector Machine (SVM) model is a supervised algorithm that creates a hyperplane which separates data into classes. The objective is to find an optimal plane that maximizes the distance between the separating hyperplane and the closed points (defined as support vectors) of the training data set. If the data is non-linear separable, there is a modified version of SVM which projects the original data to a high-dimensional space by the implementations of kernel functions. In the literature, there have been proposed different kernels such as linear, Gaussian and polynomial. In the case of a multi-class scenario, the SVM model assigns the label of +1 to one of them and -1 to all the remaining classes. This results in K binary SVM models, hence a model for each k class. This strategy is known the multi-class approach one versus all, and based on the principle of the "winner-takes-all".

3 Database

The kitchen is one of the home's spaces where different sound sources can occur at the same time specially when cooking. For this work we are interested in a kitchen environment where three different sound sources are occurring at the same moment. We believe that by mixing three sounds it can achieve a kitchen environment more realistic. Sounds mixing process considers as background disturbance (the noise) two of the three sounds sources, and the remain sound is the acoustic event (the signal) to be recognized. Additionally, we add an extra component to the sounds mixing process, which it consists of making the identification of the signal into the noise more perceptually difficult. The previous can be carried out using 3dB (decibels) of SNR.

In the literature, it is common to find databases containing different kinds of acoustic events, however, it is difficult to find a database with a mixture of kitchen sounds. Due to the above, our work consisted in building a database using the scheme presented in Beltrán-Márquez et al [5]. Sixteen archives of audio were collected where each one is a class of kitchen sound. The portals where these sounds were downloaded are, `www.soundsnap.com`, `www.freesfx.co.uk` and `www.sounddogs.com`. The audio files are WAV format, with a 44100Hz of sampling frequency and coded to 16 bits. No copyright infringement was intended. The downloaded sounds are presented in Table 2.

The audio signatures approach suggests the use of signatures between one to fifteen seconds. All downloaded audio files have a length of three seconds (we consider that three seconds of audio is enough to identify a sound from the environment). As indicated above, the database

consists of sixteen original sounds for mixing. First, mixing process consists of forming a dataset with the mixture of all the combinations of pairs of sounds. Second, all the elements from the dataset are combined with each one of the sixteen original sounds for getting mixtures with three sounds. Repetitions of sounds in a single mixture are avoided. All mixtures are obtained using 3dB of SNR, for this, the sixteen original sounds are considered the "signal" (the acoustic events to identify) and the elements of the dataset as the "noise". Figure 3 shows a illustration of the mixing process of sounds. The equation $SNR = 10 \times log_{10}(P_{signal}/P_{noise})$ is used to determine SNR between signal and noise, where P_{signal} is the power of the signal and P_{noise} is the power of the noise. Finally, the database has 1680 audio files, all of them grouped into 16 classes, where each class has 105 audio files.

In the experiments, we used classifiers such as Similarity Distance, k- Nearest Neighbors (KNN), SVM and ANN. For the experiments with KNN, ANN and SVM, we generated a training dataset to train the models of classification (this is because the elements of the database will be used as test elements to assess the classification models). This training dataset is built by using the original signal of each one of the sixteen kitchen sounds and two degraded versions of each one of them (this procedure guarantees having more data for training since there are not more instances for each class of sound). Degradation consists of distorting the signal by adding white Gaussian noise. We use $awgn(signal, SNR)$ MATLAB® function for this matter, where $signal$ is the original kitchen sound and SNR take the value of 35dB and 50dB respectively for each degraded version. The total number of audios in the training dataset is of 48. Original and mixtures of audios are available in `https://drive.google.com/open?id=1ALkT-nVt3HMFK66CjcWrC3dHrNhiyZuk`

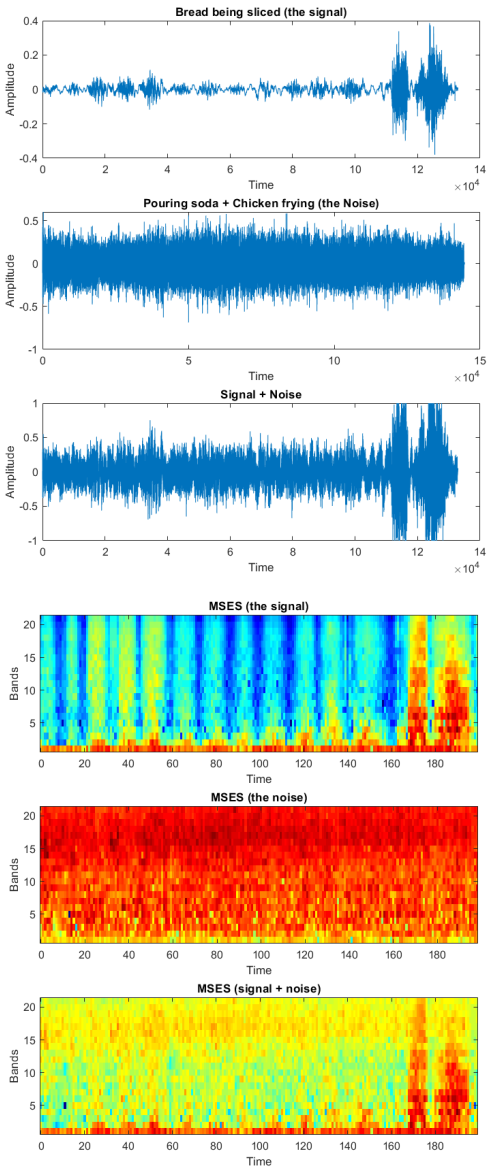

Fig. 3. Illustration of the mixing process of sounds considering the acoustic event named "Bread being sliced" as the signal and the couple of sounds "Pouring Soda - Chicken Frying" as the noise. First three panels show the signals in time domain and the last panels show the MSES signatures associated to every signal

4 Experiments

In this work, we use measures of similarity as baseline experiment to have a starting point or a first measurement in relation to the performance indicators of the considered classifiers. Certainly, the search by similarity identifies which candidate identities are more similar to one or more input entities for coincidence. In the next section,

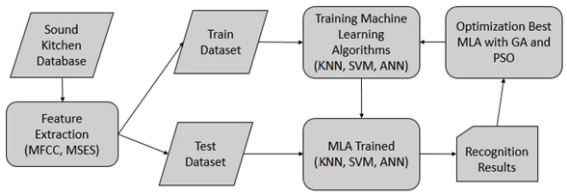

Fig. 4. Flow chart of the activities of the proposed approach

we describe how compute this entities from an approach of audio signatures.

On the other hand, as previously stated, the ANN and SVM algorithms have already been used in acoustic event classification tasks [7, 14, 24, 29, 51, 36]. Therefore, we consider it appropriate to include these models in our experiments by using the Bayesian optimization strategy for SVM and different architectures for ANN to identify their best parameters. Also, as a baseline model, the KNN algorithm was included in our study because of its easy implementation, and for this particular case, to test different distance metrics and number of neighbors.

To understand the process to be followed in our experiments, Figure 4 shows the block diagram of the sequential activities carried out in this section.

4.1 MFCC and MSES Signatures

To extract both MFCC and MSES signatures, the next procedure was implemented. a) First, stereo signals are changed to monoaural by averaging both channels, and each audio is cut to three seconds of length. b) Frames of 30ms are used to divide the monoaural signal (i.e. we use 1323 samples per frame using a sampling frequency of 44100Hz). c) Consecutive frames have an overlap of 50%, hence, there are 200 frames (T = 200) for three seconds of audio. d) A Hann window function is applied to each frame. e) The FFT is computed for each frame.

With the FFTs we are ready to compute Mel Frequency Cepstral Coeffcients (section 2.1), and the Multiband Spectral Entropy Signature (section 2.3). An additional point is that MSES signatures are extracted considering a bandwidth of 0Hz up to 8000Hz, hence, only 21 critical bands are used. The above entails each feature vector be 21-dimensional (L = 21). To have similar conditions between MSES and MFCC features, we compute MFCC using 21 triangular band-pass filters within the bandwidth mentioned before. Besides, MFCC vectors are also 21-dimensional.

4.2 Baseline Experiment with Similarity Distances

Baseline experiment consists of using similarity distances for recognition of acoustic events from the database of the kitchen sounds. Baseline experiment considers two different signatures, one uses normalized values and the other binary values. To normalize the signatures, we normalized the $L \times T$ matrix by computing the mean and standard deviation of all data of the matrix.

Haitsma's work presents a method to binarize audio signatures. This method consists of taking the sign of the differences between consecutive values [22]. For the baseline experiment the sign of the differences is encoded using $s_{ij} = 1$, if $v_{ij} - v_{ij-1} \geq 0$ and $s_{ij} = 0$ by other way, where s_{ij} denotes the i, jth binary value, v_{ij} denotes the ith value referred to the frame j, and v_{ij-1} denotes the ith value referred to the frame $j-1$ of the signature, for $i = 1, 2, ..., L$ and $j = 1, 2, ..., T$.

4.3 Experiment with Artificial Neural Networks

This experiment consists of training neural networks to classify the acoustic events that are considered the signal (not the noise) in the audios of the database. Two neural networks were considered, one trained with MFCC signatures and the other trained with MSES signatures. To train the neural networks, we used the normalized signatures that are extracted from each audio of the training dataset. Therefore, we have 48 signatures for training the neural network for MFCC and other 48 signatures for training the neural network for MSES.

For both MFCC and MSES, the neural networks consist of 2 hidden layers and 16 neurons in the output layer; the input layer has 4200 neurons (i.e., each signature of size 21×200 is converted to vector). The design of each ANN was proposed

according to the works cited in [18, 31] for the selection of the different elements of the learning algorithms. We tested three designs of neural networks with the following architectures: In the first design, the descendent gradient with adaptive learning rate back-propagation is implemented with 95 neurons in the first hidden layer and 28 neurons in the second hidden layer. For the second design, the descendent gradient with momentum and adaptive learning rate back-propagation is utilized with 150 neurons in the first hidden layer and 35 neurons in the second hidden layer.

In the third design, the scaled conjugate gradient back-propagation is applied with 79 neurons in the first hidden layer and 22 neurons in the second hidden layer. To set the number of neurons, a search was made for the best performance of the neural network in the learning stage by increasing one neuron from 10 up to 200 in the hidden layers. Finally, for each ANN, the first and second hidden layers, the hyperbolic tangent sigmoid transfer function is applied and for the output layer, the logarithmic-sigmoid transfer function is implemented.

The classification process consisted of assessing the neural networks using the normalized signature of the mixture of kitchen sounds of the database. If a neural network correctly classifies a given acoustic event in the entire database, then there will be 105 true positives for that class. The performance goal and numbers of epochs for all the neural networks are 1e-06 and 8000, respectively.

4.4 Experiment with Support Vector Machines

The same training dataset used for the ANN is used for the experiments with SVM. In our implementation, the $fitcsvm()$ MATLAB® built-in function has been used to train the SVM classifiers. There were trained 16 binary SVM models, one for each kitchen sound class. Gaussian, linear and polynomial kernels were compared in order to select the most appropriate for each model. The Bayesian optimization strategy was implemented in order to select optimal hyper-parameters by the evaluation of 30 models for each binary classifier. The best results were achieved with Gaussian kernels and the Sequential Minimal Optimization solver. Once the parameters of the 16 SVM models were defined, each mixture of sounds is classified with the model that achieved the highest score.

4.5 Experiment with K-Nearest Neighbors

Similar than the models based on SVM, the optimizer hyper-parameter function of MATLAB®, $fitcknn()$, was implemented to perform a Bayesian optimization strategy. In this implementation, different distance metrics, such as Euclidean, City-block, Cosine, Minkowski, Correlation, Spearman, Hamming, Mahalanobis, Jaccard, and Chebychev, were evaluated. Also, different number of neighbors were implemented within each search. In total, there were compared the performance of 30 different models.

5 Results and Discussion

In this section, we compare results about the performance of MSES and MFCC using four types of classifiers: similarity distances, KNN, ANN and SVM. Results are showed using True Positives (TP) and False Positives (FP) from the confusion matrices, the best experimental outcomes and the averages achieved with each classifier are summarized in Table 7.

5.1 Similarity Distance Results

Table 2 shows the results for each signature using Hamming distance and Cosine distance, here the recall metric is used for results comparison. Although it is common to use binary signatures in an audio signature-based approach, the results of Table 2 suggests that binary signatures are not convenient to represent acoustic events, especially, when they have non-stationary characteristics.

The difference in recall between both features is about the 3.46%, therefore, no advantage can be seen by using MFCC or MSES features. An audio signature using normalized values seens to work better, allowing to differentiate more the performance of both feature extraction methods, especially, when working with low levels of SNR.

Table 2. Results about Recall

Sounds[a]	Hamming Distance		Cosine Distance	
	MFCC	MSES	MFCC	MSES
C1	43.80	**51.42**	49.52	**95.23**
C2	**0**	**0**	0.95	**6.66**
C3	**100**	**100**	90.47	**94.28**
C4	**100**	**100**	84.76	**100**
C5	99.04	**100**	80	**80.95**
C6	**100**	98.09	**100**	51.42
C7	**41.90**	40	44.76	**97.14**
C8	**65.71**	62.85	42.85	**100**
C9	27.61	**40**	36.19	**45.71**
C10	57.14	**79.04**	69.52	**78.09**
C11	14.28	**38.09**	**28.57**	17.14
C12	**43.80**	35.23	61.90	**87.61**
C13	**100**	**100**	94.28	**96.19**
C14	53.33	**70.47**	63.80	**97.14**
C15	**98.09**	84.76	85.71	**100**
C16	**100**	**100**	81.90	**100**
Average	65.29	**68.75**	63.45	**77.97**

[a] The different acoustic events are: (C1) Bread being sliced, (C2) Chop food quickly and strongly, (C3) Pouring soda into a glass, (C4) Electric blender liquefying food, (C5) Frying chicken in a pan, (C6) Hot oil in a pan, (C7) Burner of a stove, (C8) Making popcorn in a microwave, (C9) Cooking fryer, (C10) Peeling potatoes, (C11) Making popcorn in a pot, (C12) Turning a microwave on and off, (C13) Pouring water into a glass, (C14) Slicing onions, (C15) Boiling teapot, and (C16) Boiling eggs.

Hamming distance results, Table 2, shows that C2 was the worst classified class because it has zero in recall score, while, C3, C4, C13 and C16 are the classes of sounds with the higher recall in both features, 100% in all of them. The average recall for MFCC features is 65.29% and 68.75% for using MSES features.

The results with Cosine distance using MSES feature, shows that C4, C8, C15 and C16 are the classes of sounds getting the higher recall score, whereas C2 was the worst classified class for both features.

In this case, the average recall obtained for MFCC features is 63.45% and 77.97% for MSES features (i.e., the difference of recall between both features is about the 14.52%).

The results of this experiment mark the baseline and the starting point to evaluate the contribution of machine learning methods. An image-based representation of the results with similarity distances, KNN, SVM and ANN methods for both MFCC and MSES methods is showed in Figure 5 using confusion matrices.

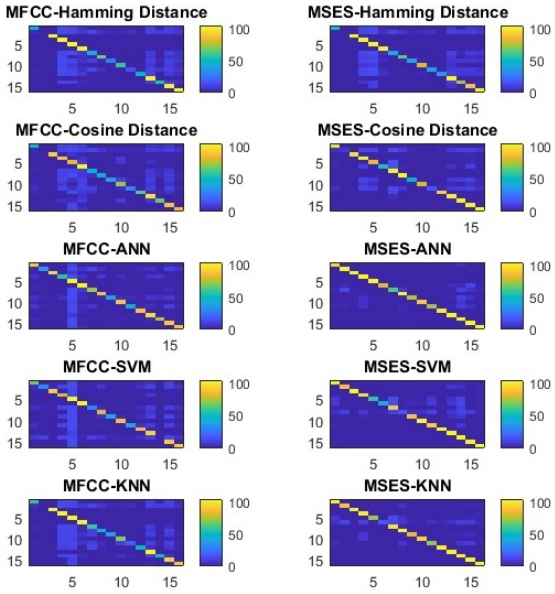

Fig. 5. Confusion matrices obtained from similarity distances, KNN, SVM and ANN methods.

5.2 Artificial Neural Network Results

Table 3 shows the results obtained with the neural network architectures using back-propagation with gradient descendent and adaptive learning rate (NNGDA), gradient descendent with momentum and adaptive learning rate (NNGDX) and scaled conjugate gradient (NNSCG). The best recall achieved for MFCC features is of 75% and for MSES features is 90.95%, both with NNGDX. The average is obtained for 30 experiments, but only 10 experiments are presented in Table 3. The best average recall score was 73.42% and 88% for MFCC and MSES respectively, both from NNGDX method.

Table 4 shows the results about True Positives (TP) and False Positives (FP) from the confusion matrix obtained for the best performance with artificial neural networks using MFCC and MSES, respectively. For MFCC, C5 and C6 are the classes that obtained the higher scores, 1 and 2 errors

Table 3. Results for artificial neural networks using recall metric

Experiment	NNGDA		NNGDX		NNSCG	
	MFCC	MSES	MFCC	MSES	MFCC	MSES
1	73.21	88.1	**75**	**90.95**	73.69	89.05
2	73.15	87.92	**74.88**	**90.24**	73.51	88.99
3	73.04	87.8	**74.52**	86.76	73.27	**88.33**
4	73.04	87.8	**74.17**	**89.23**	73.15	88.21
5	72.98	87.8	**74.11**	**89.17**	73.15	88.21
6	72.92	87.68	**73.87**	**89.17**	73.1	88.1
7	72.92	87.5	**73.81**	**89.11**	72.98	87.74
8	72.86	87.5	**73.81**	**89.05**	72.92	87.74
9	72.8	87.5	**73.75**	**88.75**	72.92	87.62
10	72.8	87.44	**73.75**	**88.69**	72.86	87.62
Best Result	73.21	88.1	**75.00**	**90.95**	73.69	89.05
Average	72.62	86.94	**73.42**	**88.00**	72.59	87.23

Table 4. Results about True Positives (TP) and False Positives (FP) with artificial neural networks

Sounds	MFCC		MSES	
	TP	FP	TP	FP
C1	91	32	98	5
C2	37	5	**105**	18
C3	97	3	104	0
C4	56	5	**105**	17
C5	104	176	**105**	1
C6	103	41	81	1
C7	72	19	63	0
C8	88	18	104	14
C9	38	0	77	1
C10	89	15	90	0
C11	46	4	75	1
C12	89	9	101	0
C13	96	26	**105**	9
C14	75	16	**105**	48
C15	90	37	**105**	31
C16	89	14	**105**	6

respectively. At least 14 samples of each class of kitchen sounds (except by C6) are classified erroneously as C5.

For MSES, C7 is the class with more errors, 42 in total, followed by C11 (30 errors) and C9 (28 errors). Unlike the experiment with similarity distances, here the sound class C2 is 100% classified. The others classes with higher scores are C3, C4, C5, C8, C13, C14, C15 and C16. Indeed, experiments with ANN show that there is an increase in the recall with which kitchen sounds are identified. Comparing the average value achieved with distances of similarity and neural networks, there is an increase of recall of 8.13% for MFCC and 10.03% for MSES.

5.3 Support Vector Machine Results

Table 5 shows the results for TP and FP from the confusion matrix obtained with the SVM classifier using MFCC and MSES features, respectively. The recall obtained by using the MFCC features is 67.2%. C5 and C6 are the classes that obtained the higher recall, zero and one errors respectively. For MFCC, at least 14 samples of each class of kitchen sounds (except by C5 and C6) are classified erroneously as C5. All the sound samples of C14 are miss classified (105 errors). C7 and C2 obtained 77 and 73 errors, respectively. For MSES, the recall achieved is 83.99%. C8 is the class with more errors, 89 in total, followed by C6 (61 errors) and C5 (46 errors). Comparing the average value achieved with distances of similarity and SVM, there is an increase of recall of 1.91% for MFCC and 6.02% for MSES.

5.4 K-Nearest Neighbors Results

As previously mentioned, the $fitcknn()$ MATLAB® function was used to compare the performance of 30 different models. The one that obtained the best performance with MFCC features was the model that uses the Spearman distance function with two neighbors. Table 6 shows the results about TP and FP from the confusion matrix of this implementation. The recall metric was 65.77%. C3, C4, C5, C6 and C13 obtained the best results. Contrary, only one sample of C2 was correctly classified. The results obtained with MSES feature (Table 6) showed that the best KNN model uses the correlation distance function and one neighbor. The recall metric obtained was 87.38%. C8, C13, C14 and C16 obtained zero errors in classification. Four classes obtained between 1, 2 or 3 errors. The more difficult class to identify was C6 with 88 errors in total.

Comparing the average value achieved with distances of similarity and KNN, there is an increase of recall of 0.48% for MFCC and 9.41% for MSES.

Table 7 shows the summary of the obtained results for both MFCC and MSES features and all classifiers: Similarity distances, ANN, SVM and KNN. We can observe first that all the classifiers

Table 5. Results about True Positives (TP) and False Positives (FP) with support vector machine

Sounds	MFCC		MSES	
	TP	FP	TP	FP
C1	68	24	103	26
C2	32	13	88	12
C3	93	3	104	0
C4	81	63	97	10
C5	**105**	191	59	0
C6	104	3	44	2
C7	28	10	90	40
C8	83	29	16	0
C9	43	4	89	9
C10	88	9	**105**	9
C11	40	3	93	27
C12	88	31	**105**	7
C13	94	72	**105**	31
C14	0	0	**105**	71
C15	91	72	103	9
C16	91	24	**105**	16

Table 6. Results considering True Positives (TP) and False Positives (FP) for k-nearest neighbors

Sounds	MFCC		MSES	
	TP	FP	TP	FP
C1	58	11	104	18
C2	1	0	90	12
C3	104	1	103	0
C4	**105**	113	103	25
C5	104	134	73	1
C6	**105**	37	17	0
C7	60	57	100	29
C8	45	17	**105**	34
C9	39	3	85	3
C10	72	21	95	0
C11	19	6	72	5
C12	50	2	102	4
C13	**105**	77	**105**	11
C14	54	9	**105**	22
C15	94	70	104	29
C16	90	17	**105**	19

have an improvement in the recall metric when working with MSES feature.

Second, the ANN classifier has the highest performance for both MFCC and MSES (73.42% and 88%, respectively), followed by a combination MSES-KNN (87.38%), then a combination MSES-SVM (83.99%), and finally, similarity distances-MSES with a score of 77.97%. Regarding MFCC, the second best performance was achieved with SVM (i.e., 67.2%).

Table 7. Best results for the classification of kitchen sounds

Method	Feature	
	MFCC (%)	MSES (%)
Similarity Distance	65.29	77.97
ANN	**73.42**	**88.00**
SVM	67.20	83.99
KNN	65.77	87.38

Third and fourth best performance were achieved with KNN and similarity distances (65.77% and 65.29%, respectively). We attribute the good performance of ANN to the fact this machine learning technique works with variations that allow their learning to be more robust and effective than the other methods.

5.5 Test of Statistical Significance

To further analyze the differences between MFCC and MSES methods, we applied a non-parametric Mann-Whitney's test with a significance level of $\alpha = 0.05$. For this test, two population samples were related which belong to the recall metric scores of the 30 models evaluated using ANN, SVM and KNN for both MFCC and MSES features. The results show a value $p = 0.0003$, which makes us reject the null hypothesis and conclude that the medians of both methods are different and that they do not depend on the type of classifier or the sounds to be recognized.

5.6 Optimization of ANN with GA and PSO

Previous results showed that the combination MSES-ANN (audio features-classifier) achieved the best score for all the combinations. In this part, we realized an optimization looking for the best artificial neural network with MSES. This optimization is performed using the genetic algorithm (GA) and particle swarm optimization (PSO). The use of the GA and PSO optimization algorithms are decided in consideration because

these algorithms performed good results in optimization of parameters for machine learning algorithms [19, 20].

The optimization looks up for the following ANN's values and parameters:

1. Number of neurons in the first hidden layer.
2. Number of neurons in the second hidden layer.
3. The transfer functions for the neurons in the first and second hidden layer, and for the neurons in the output layer. The transfer functions for optimizing are the next:
   - Positive linear transfer function.
   - Linear transfer function.
   - Inverse transfer function.
   - Log-sigmoid transfer function.
   - Hyperbolic tangent sigmoid transfer function.
   - Triangular basis transfer function.
   - Hard-limit transfer function.
   - Saturating linear transfer function.
   - Elliot symmetric sigmoid transfer function.
   - Symmetric saturating linear transfer function.
   - Symmetric hard-limit transfer function.
   - Elliot 2 symmetric sigmoid transfer function.
4. The learning algorithms implemented in the neural network:
   - Levenberg-Marquardt backpropagation.
   - One-step secant backpropagation.
   - Gradient descent with adaptive learning rate backpropagation.
   - Gradient descent with momentum and adaptive learning rate backpropagation.
   - Scaled conjugate gradient backpropagation.
   - Resilient backpropagation.
   - Gradient descent backpropagation.
   - Gradient descent with momentum backpropagation.
   - Conjugate gradient backpropagation with Fletcher-Reeves updates.
   - Conjugate gradient backpropagation with Polak-Ribiére updates.
   - Conjugate gradient backpropagation with Powell-Beale restarts.

Table 8. Parameters for GA

Population	100 Individuals
Individual	6 Genes (real)
Generations	100
Assign Fitness	Ranking
Selection	Stochastic universal sampling
Mutation	16.67 %
Crossover	Single Point (80%)

Table 9. Parameters for PSO

Population	100 Particles
Particle	6 Dimensions (real)
Iterations	100
Constriction Coefficient	1
Inertia Weight	0.1
R1, R2	Random in the range [0,1]
C1	Lineal decrement (2–0.5)
C2	Lineal increment (0.5–2)

In Table 8, the parameters for the performance of GA are showed and Table 9 shows the parameters for the performance of PSO.

Table 10 shows the results for acoustic event recognition for 10 experiments that combine MSES-ANN with both optimization techniques GA and PSO. The average in recall metric was 91.46% and 91.55% for GA and PSO, respectively,

The best recall for the optimization of the neural network was obtained with PSO achieving a 93.93 % of recognition for the kitchen sounds. The parameters of the best ANN architecture with PSO are:

- 1st Hidden layer (1HL) with 186 neurons.
- 2nd Hidden layer (2HL) with 238 neurons.
- The transfer function in 1HL was saturating linear transfer function.

Table 10. Results about optimization of ANN-MSES with GA and PSO

Experiment	Algorithm	
	GA	PSO
1	92.20	90.89
2	91.31	89.35
3	91.67	91.85
4	91.01	93.21
5	91.67	91.90
6	91.31	91.37
7	91.19	91.13
8	91.13	90.71
9	91.19	91.13
10	91.90	93.93
Best result	**92.20**	**93.93**
Average	**91.46**	**91.55**

Table 11. Results with the best ANN architecture

Experiment	**Recall**
1	94.52
2	93.51
3	94.11
4	94.70
5	93.21
6	93.39
7	93.15
8	93.81
9	93.04
10	93.10
Best result	**94.70**
Average	**93.46**

— The transfer function in 2HL was symmetric saturating linear transfer function.

— The transfer function in output Layer was symmetric saturating linear transfer function.

— The training learning algorithm was conjugate gradient backpropagation with Fletcher-Reeves updates.

Finally, 30 experiments were realized using ANN with the above configuration parameters with the aim of testing the optimization robustness. Table 11, presents only the best 10 results where one can observe that the average recall achieved for the optimized combination MSES-ANN was 93.46%, that is, 15.49% of improvement when compared with the average value achieved with distances of similarity and optimized ANN-MSES.

Table 12 shows the results about True Positives (TP) and False Positives (FP) from the confusion matrix obtained for the best performance with the couple MSES-ANN and optimized with PSO. The results of the table showed that, excepting the C7 sound, all classes have a success ratio between 90% and 100% for the recognition of acoustic events that define each class. The optimization of the neural network helps to improve the recognition rate and to reduce the number of miss classified sounds (False Positives). An image-based representation of the confusion matrix of this experiment is showed in Figure 6. Notice that the color of the diagonal indicates that there is a high recognition rate for each of the classes.

6 Conclusions

In this work, we identify acoustic events using the approach of audio signatures in combination with machine learning algorithms. When different instances of a sound class are not available, the audio signatures approach should be used since this approach only requires the original sound and degraded versions of it.

Audio signatures help us to cope with the small database of the kitchen sound sources, which in our case consisted of sixteen original sounds and some degraded versions of these. In order to complement the audio signatures approach, we studied the performance of machine learning algorithms when there is only an instance of the sound classes and degraded versions of them.

The two audio features considered in this work are MFCC and MSES. MFCC is one most cited audio feature when working with audio-based activity recognition, and the reason to be considered as our benchmark feature. MSES is an interesting audio feature being widely adopted because of its robustness to noise.

The results showed that the representation of acoustic events based on MSES is more convenient when working with different classification methods. Although the comparison between MSES and MFCC is not conclusive, it seems that

Table 12. Results about True Positives (TP) and False Positives (FP) with optimized ANN

Sounds	MSES	
	TP	FP
C1	96	1
C2	99	6
C3	104	0
C4	103	14
C5	104	5
C6	100	1
C7	63	0
C8	104	3
C9	102	4
C10	95	0
C11	102	6
C12	103	2
C13	104	3
C14	104	16
C15	103	3
C16	**105**	26

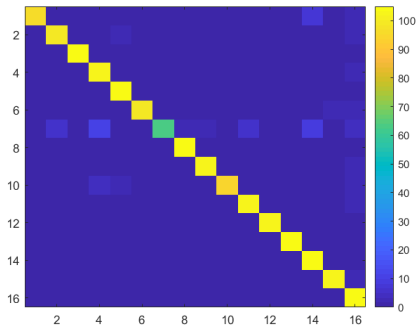

Fig. 6. Confusion matrix obtained with the couple ANN-MSES and optimized with PSO

MSES is an audio feature that is very robust for identifying acoustic events in a mixture of sounds.

One thing to note is that MSES captures the location of energy peaks in each sub-band that are less corrupted by noise, even in the presence of low SNR levels, something that affect the performance of MFCC. Nevertheless, both MFCC and MSES represent very well the non-stationary characteristics of audio signals.

A database with a mixture of everyday kitchen sounds was created using 3dB of SNR. The way in which this database is constructed should encourage readers to use it in future works since this database considers noisy contexts, something that to our understanding is not available in the literature. Yet there are databases with sound sources from different and independent tasks but never mixed, such the one provided by the DCASE2020 database.

The results presented here showed a way for identifying acoustic events when they are immersed in a mixture of sounds and they are not predominant, which is important for recognizing activities in real indoor environments. In the classification stage, four types of classifiers were used, Similarity Distances, k-Nearest Neighbors, Support Vector Machines and Artificial Neural Networks.

The results of Table 7 showed that MSES combined with Artificial Neural Networks has an score of 88% in recall metric which outperforms any other combination of classifiers with MSES or MFCC. In addition, a test of statistical significance was realized, getting a value of p = 0.0003, which makes us reject the null hypothesis and conclude that MFCC and MSES features have different level of robustness and that their performance do not depend on the type of classifier nor on the sound to be recognized.

Furthermore, the use of a genetic algorithm and a particle swam optimization improved the performance of audio features recognition supported by machine learning classifiers, being the combination MSES-ANN the one the best performance (93.46%). Table 10 showed that PSO performs better than GA achieving a average recall of 91.55%.

Finally, the experiments presented in this work focused on the evaluation MSES and MFCC audio features techniques that are supported by machine learning algorithms for the recognition of acoustic events on noisy environments. We considered the context of a kitchen context where different sound sources are present, for instance, when a person is preparing meals. In an attempt to make a more realistic scenario sound sources were mixed and applied a low SNR level. This is an acoustic recognition approach that would help better understand the nature of human activity in the home setting.

The identification of all the sounds that are present in the environment might help to develop systems that can assist people or that can be aware of potential dangers.

References

1. **Almaadeed, N., Asim, M., Al-Maadeed, S., Bouridane, A., Beghdadi, A. (2018).** Automatic detection and classification of audio events for road surveillance applications. Sensors, Vol. 18(6), pp. 1–19.
2. **Alsina-Pagès, R. M., Navarro, J., Alías, F., Hervás, M. (2017).** homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring. Sensors, Vol. 17(4), pp. 1–22.
3. **Aucouturier, J.-J., Defreville, B., Pachet, F. (2007).** The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. The Journal of the Acoustical Society of America, Vol. 122(2), pp. 881–891.
4. **Bansal, R., Shukla, N., Goyal, M., Kumar, D. (2020).** Information and Communication Technology for Intelligent Systems, chapter Enhancement and Comparative Analysis of Environmental Sound Classification Using MFCC and Empirical Mode Decomposition. Springer, Singapore, pp. 227–235.
5. **Beltrán, J., Chávez, E., Favela, J. (2012).** Environmental sound recognition by measuring significant changes in the spectral entropy. Lecture Notes in Computer Science, Mexican Conference on Pattern Recognition, Vol. 1, pp. 334–343.
6. **Beltrán, J., Chávez, E., Favela, J. (2015).** Scalable identification of mixed environmental sounds, recorded from heterogeneous sources. Pattern Recognition Letters, Vol. 68(1), pp. 153–160.
7. **Bountourakis, V., Vrysis, L., Konstantoudakis, K., Vryzas, N. (2019).** An enhanced temporal feature integration method for environmental sound recognition. Acoustics, Vol. 1(2), pp. 410–422.
8. **Bryan-Kinns, N. (2017).** Interaction design with audio: Speculating on sound in future design education. The 4th Central China International Design Science Seminar 2017, pp. 1–9.
9. **Camarena-Ibarrola, A., Chávez, E. (2006).** On musical performances identification, entropy and string matching. 2006 Mexican International Conference on Artificial Intelligence, pp. 952–962.
10. **Camarena-Ibarrola, A., Chávez, E. (2010).** Real time tracking of musical performances. 2010 Mexican International Conference on Artificial Intelligence, pp. 138–148.
11. **Camarena-Ibarrola, A., Figueroa, K., García, J. (2020).** Speaker identification using entropygrams and convolutional neural networks. 2020 Mexican International Conference on Artificial Intelligence, pp. 23–34.
12. **Camarena-Ibarrola, A., Luque, F., Chávez, E. (2017).** Speaker identification through spectral entropy analysis. 2017 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), pp. 1–6.
13. **Chachada, S., Kuo, J. (2014).** Environmental sound recognition: A survey. APSIPA Transactions on Signal and Information Processing, Vol. 3, pp. 1–15.
14. **Chandrakala, S., Jayalakshmi, S. L. (2019).** Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies. ACM Computing Surveys, Vol. 52(3), pp. 1–34.
15. **Cheng, C.-F., Rashidi, A., Davenport, M. A., Andersona, D. V. (2017).** Activity analysis of construction equipment using audio signals and support vector machines. Automation in Construction, Vol. 81, pp. 240–253.
16. **Deepsheka, G., Kheerthana, R., Mourina, M., Bharathi, B. (2020).** Recurrent neural network based music recognition using audio fingerprinting. 2020 Third International Conference on Smart Systems and Inventive Technology, pp. 1–6.
17. **Gan, G., Ma, C., Wu, J. (2007).** In Data Clustering: Theory, Algorithms and Applications, chapter Similarity and Dissimilarity Measures. ASA-SIAM Series on Statistics and Applied Probability, pp. 67–106.
18. **Gaxiola, F., Melin, P., Valdez, F., Castillo, O. (2011).** Modular neural networks with type-2 fuzzy integration for pattern recognition of iris biometric measure. Batyrshin I., Sidorov G. (eds) Advances in Soft Computing. MICAI 2011. Lecture Notes in Computer Science, Vol. 7095.

19. **Gaxiola, F., Melin, P., Valdez, F., Castro, J. (2018).** Optimization of deep neural network for recognition with human iris biometric measure. Melin P., Castillo O., Kacprzyk J., Reformat M., Melek W. (eds) Fuzzy Logic in Intelligent System Design. NAFIPS 2017. Advances in Intelligent Systems and Computing, Vol. 648.

20. **Gaxiola, F., Melin, P., Valdez, F., Castro, J., Manzo-Martínez, A. (2019).** Pso with dynamic adaptation of parameters for optimization in neural networks with interval type-2 fuzzy numbers weights. Axioms, Vol. 8(1).

21. **Grama, L., Rusu, C. (2019).** Extending assisted audio capabilities of tiago service robot. 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), IEEE Xplore, pp. 1–8.

22. **Haitsma, J., Kalker, T. (2002).** A highly robust audio fingerprinting system. Proceedings of International Symposium on Music Information Retrieval, pp. 1–9.

23. **Huang, W., Zhang, Y. (2020).** Application of hidden markov chain and artificial neural networks in music recognition and classification. Proceedings of 2020 the 6th International Conference on Computing and Data Engineering, pp. 49–53.

24. **Jatturas, C., Chokkoedsakul, S., Na-Ayudhya, P. D., Pankaew, S., Sopavanit, C., Asdornwised, W. (2019).** Recurrent neural networks for environmental sound recognition using scikit-learn and tensorflow. 2019 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, pp. 1–6.

25. **Kar, B., Samanta, S., Prasad-Manna, R., Chatterjee, S. (2019).** An optimized music recognition system using mel-frequency cepstral coefficient (mfcc) and vector quantization (vq). Special Issue International Business Research Conference on Transformation Opportunities and Sustainability Challenges in Technology and Management., Vol. 45489(1208), pp. 100–106.

26. **Kaur, G., Srivastava, M., Kumar, A. (2018).** Genetic algorithm for combined speaker and speech recognition using deep neural networks. Journal of Telecommunications and Information Technology, Vol. 2, pp. 23–31.

27. **Kumar, A. P., Roy, R., Rawat, S., Sudhakaran, P. (2017).** Continuous telugu speech recognition through combined feature extraction by mfcc and dwpd using hmm based dnn techniques. International Journal of Pure and Applied Mathematics, Vol. 114(11), pp. 187–197.

28. **Kumar, A. S., Erler, R., Kowerko, D. (2019).** Audio-based event recognition system for smart homes. Proceedings of the 27th ACM International Conference on Multimedia, ACM, pp. 2205–2207.

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

30. **Luque-Suarez, F., Camarena-Ibarrola, A., Chávez, E. (2019).** Efficient speaker identification using spectral entropy. Multimedia Tools and Applications, Vol. 78, pp. 16803–16815.

31. **Melin, P. (2012).** Modular neural networks for person recognition using the contour segmentation of the human iris. Modular Neural Networks and Type-2 Fuzzy Systems for Pattern Recognition. Studies in Computational Intelligence, Vol. 389.

32. **Misra, H., Ikbal, S., Bourlard, H., Hermansky, H. (2004).** Spectral entropy based feature for robust asr. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1–8.

33. **Misra, H., Ikbal, S., Sivadas, S., Bourlard, H. (2005).** Multi-resolution spectral entropy feature for robust asr. 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1–9.

34. **Mohammad-Djafari, A. (2001).** Entropie en traitement du signal. Laboratoire des Signaux et Systemes, pp. 1–9.

35. **Moreaux, M., Garcia-Ortiz, M., Ferrané, I., Lerasle, F. (2019).** Benchmark for kitchen20, a daily life dataset for audio-based human action recognition. 2019 International Conference on Content-Based Multimedia Indexing, pp. 1–6.

36. **Mushtaq, Z., Su, S.-F. (2020).** Environmental sound classification using a regularized deep convolutional neural network with data augmentation. Applied Acoustics, Vol. 167, pp. 1–13.

37. **Naithani, K., Thakkar, V. M., Semwal, A. (2018).** English language speech recognition using mfcc and hmm. 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE), IEEE Xplore, pp. 1–7.

38. Naronglerdrit, P., Mporas, I. (2017). Interactive Collaborative Robotics, chapter Recognition of Indoors Activity Sounds for Robot-Based Home Monitoring in Assisted Living Environments. Springer, Cham, pp. 153–161.

39. Naronglerdrit, P., Mporas, I., Sotudeh, R. (2017). Improved automatic keyword extraction given more linguistic knowledge. 2017 IEEE 13th International Colloquium on Signal Processing and its Applications (CSPA), IEEE Xplore, pp. 23–28.

40. Pires, I. M., Santos, R., Pombo, N., Garcia, N. M., Flórez-Revuelta, F., Spinsante, S., Goleva, R., Zdravevski, E. (2018). Recognition of activities of daily living based on environmental analyses using audio fingerprinting techniques: A systematic review. Sensors, Vol. 18(1), pp. 1–23.

41. Ren, F., Bao, Y. (2020). A review on human-computer interaction and intelligent robots. International Journal of Information Technology and Decision Making, Vol. 19(1), pp. 5–47.

42. Robinson, F. A., Bown, O., Velonaki, M. (2020). Implicit communication through distributed sound design: Exploring a new modality in human-robot interaction. Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, ACM, pp. 597–599.

43. Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, Vol. 27(3), pp. 379–423.

44. Shen, J.-l., Hung, J.-w., Lee, L.-s. (1998). Robust entropy-based endpoint detection for speech recognition in noisy environments. 5th International Conference on Spoken Language Processing, pp. 1–4.

45. Shen, Y.-H., He, K.-X., Zhang, W.-Q. (2018). Home activity monitoring based on gated convolutional neural networks and system fusion. DCASE2018 Challenge Tech. Rep., pp. 1–5.

46. Sigurdsson, S., Petersen, K. B., Lehn-Schiøler, T. (2006). Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. 2006 International Society for Music Information Retrieval, pp. 1–4.

47. Smith, J. O., Abel, J. S. (1999). Bark and erb bilinear transforms. IEEE Transactions on Speech and Audio Processing, Vol. 7(6), pp. 697–708.

48. Telembici, T., Grama, L., Rusu, C. (2020). Integrating service robots into everyday life based on audio capabilities. 2020 International Symposium on Electronics and Telecommunications (ISETC), IEEE Xplore, pp. 1–8.

49. Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. The Journal of the Acoustical Society of America, Vol. 88(97), pp. 97–100.

50. Vafeiadis, A., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., Hamzaoui, R. (2017). Audio-based event recognition system for smart homes. 2017 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, IEEE Xplore, pp. 1–8.

51. Wei, P., He, F., Li, L., Li, J. (2020). Research on sound classification based on svm. Neural Computing and Applications, Vol. 32, pp. 1593–1607.

52. Zhang, X., Zou, Y., Shi, W. (2017). Dilated convolution neural network with leakyrelu for environmental sound classification. 2017 22nd International Conference on Digital Signal Processing, IEEE Xplore, pp. 1–5.

53. Zhang, Z., Xu, S., Cao, S., Zhang, S. (2018). Pattern Recognition and Computer Vision, chapter Deep Convolutional Neural Network with Mixup for Environmental Sound Classification. Springer, Cham, pp. 356–367.

Article received on 01/06/2021; accepted on 18/11/2021.
Corresponding author is Alain Manzo-Martínez.