Kitchen Acoustic Event Identification based on the Entropy of a Random Process

Alain Manzo-Martínez[a], Fernando Gaxiola[a], Graciela Ramírez-Alonso[a], Raymundo Cornejo[a], Luis Carlos González-Gurrola[a], Antonio Camarena-Ibarrola[b]

[a]*Facultad de Ingeniería, Universidad Autónoma de Chihuahua, Circuito Universitario Campus II, Chihuahua Chih., C.P. 31240, México.*
[b]*Facultad de Ingeniería Eléctrica, Universidad Michoacana de San Nicolás de Hidalgo, Av. Francisco J. Mújica, Morelia Mich., C.P. 58030, México.*

Abstract

The work about audio recognition has been directed to speech and music decades ago. However, the issue of classification and recognition of acoustic events has received more attention the last years. This is due to the importance of describing the context of a scenario from the analysis of different sound sources. Our work is focused on the entropy of a random process for computing the Multiband Spectral Entropy Signature (MSES) from a mixture of sounds that are produced in a kitchen environment, and the classification of these sounds. We compared our approach against a representation based on the 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. The results bear out that MSES outperforms to MFCC for getting a better score in recall metric.

Keywords: Acoustic event recognition, Audio signatures, Spectral entropy, Neural networks

1. Introduction

Acoustic events refer to several everyday sounds which are generated in natural or artificial form (i.e., 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 [30], ambient assisted living [26], human – computer interaction [13], home security surveillance [34], assisted robotics [25], between many 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. Limiting the universe of sounds may help to improve the recognition stage [3]. Recently, some research about AERS has focused on two types of classification, the first one is about acoustic events classification for a specific context and the second one is about classification of acoustic events into contextual classes [11]. For example, in a home environment, activity recognition systems using 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, if it is about activity recognition from the sounds that occur in different places, 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.

The first works about AERS were based on the paradigms of speech and music. However, the non-stationary characteristics of the acoustic events contributed to these works to be ineffective for the databases with a great number of sound sources [8]. For example, in speech recognition is common to use a phonetic structure that can be seen as a basic component of voice. This structure allows modeling complex spoken words by dividing in elemental phonemes that can be modeled by probabilistic models. Conversely, acoustic events such as a car crash or to pour some liquid have not apparent structures as the phonemes. Even, if it was possible to identify and to learn a dictionary of basic units of these events, it would be difficult for modeling its variation in time. In the same way, the comparison between music and acoustic event, the latter does not exhibit significant stationary patterns such as melody and rhythm [8].

An AERS 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 features extraction algorithms use a scheme called bag-of-frames. The bag-of-frame approach consists in considering the signal in a blind way, using a systematic

and general scheme where the signal is divided into consecutive, possibly overlapping frames, from which a vector of features is determined [1]. 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.

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 [9]. Since MFCC has been successfully used in speech and music applications, some work uses it for characterizing acoustic events containing a large and diverse variety of sounds, including those with strong temporal domain [20, 19]. In addition to the above, MFCC are often used by researchers for benchmarking their works [15].

After working the feature extraction phase, the recognition phase of an AERS can be implemented in different ways, for instance, Support Vector Machine (SVM) techniques have won an important site in recognition tasks. SVM is a classifier that discriminates the data by creating boundaries between classes rather than estimating class conditional densities. The before mentioned means that SVM may need considerably fewer data to perform accurate classification. In fact, SVM has already been used for acoustic event classification [2, 16, 32].

Nowadays, Artificial Neural Networks (ANN) with the help of new training mechanisms, it is giving promising results in many 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. Due to the similar nature of the problems in speech and music, ANN have been successful at AERS when working with large amounts of training data, acoustic event overlapped or when hierarchical features are required [24, 31, 35, 10].

There are other techniques that can be used to identify acoustic events, these techniques find the acoustic events that sound similar to the audio that the system captures. If the captured audio signals are short segments, usually this technique is known as audio signature approach and a distance function is used to identify them. Audio signatures use two fundamental processes to be determined, a feature extraction process and a modeling process; the latter refers to the representation that compactly describes a signal, so it is as robust as possible against typical audio degradations [12]. Audio signatures 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.

The approach used in this work is to treat the signals unprocessed, no source separation technique is used. 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), which it is our proposal, since this feature has never been studied to recognize acoustic events exclusively in 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, our intention is not to classify different instances of acoustic events into classes, but to evaluate different classifiers by identifying contaminated acoustic events from the sounds stored in a database.

Toward the future goal of providing assistance to elder people living alone by capable systems of identifying acoustic events of risk in a kitchen environment, we have developed a database of sounds collected from real kitchens. It should be noted that there is no database in the literature similar to ours, since 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.

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 [1]. This process of splitting the signal into frames is a characteristic part for computing MFCC and MSES, therefore, the next subsections describe the process for determining both audio features.

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 [17].

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 $w(n) = 0.5 + 0.5cos(2\pi n/N)$.

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

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),

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

for $m = 1, 2, ..., M$, where M is the number of filters and $M \ll N$. The Mel filter bank 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),

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

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 [28], this implementation considers a filter bank 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 [27]. 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.

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

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 [14], 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 [21], 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, [22] 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. [21, 22], we use spectral entropy concept for getting a robust signature that can be used in different audio recognition issues [4, 5, 18, 6, 7]. Unlike Misra et al., we compute entropy at each sub-band by using the entropy of a random process.

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 mean zero and covariance matrix, $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_x})$, then replacing $p(\mathbf{x})$ into (5), we get this known equation for determining the entropy of a vector on a random process [23], such as shown in (6), where $|\mathbf{\Sigma_x}|$ is the determinant of the covariance matrix.

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

For computing MSES, the audio signal should be divided into frames for extracting of each one a vector with L coefficients of entropy. Next, the Short Time Fourier Transform is computed on each frame by using (1). For dividing the full-band spectrum in sub-bands, we take into account the idea about how people identify sounds. The human ear perceives better the lower frequencies than the 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 [29, 33]. Table 1 shows the first 24 critical bands with their respective bandwidths.

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

Table 1: Critical bands for the Bark scale.

We use (7) to change of 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 by (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 $H = ln(2\pi)+(1/2)ln(\sigma_{xx}\sigma_{yy}-$

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

Figure 1 shows the signatures of two acoustic events that are obtained with the MSES method. In the upper panels, the signals in time domain of the acoustic events "Bread being sliced" (left) and "Microwave On-Off" (right) are showed. In the middle panels, spectrograms of both signals are showed. Finally, in the lower panels, a illustration of the signatures for both acoustic events is showed.

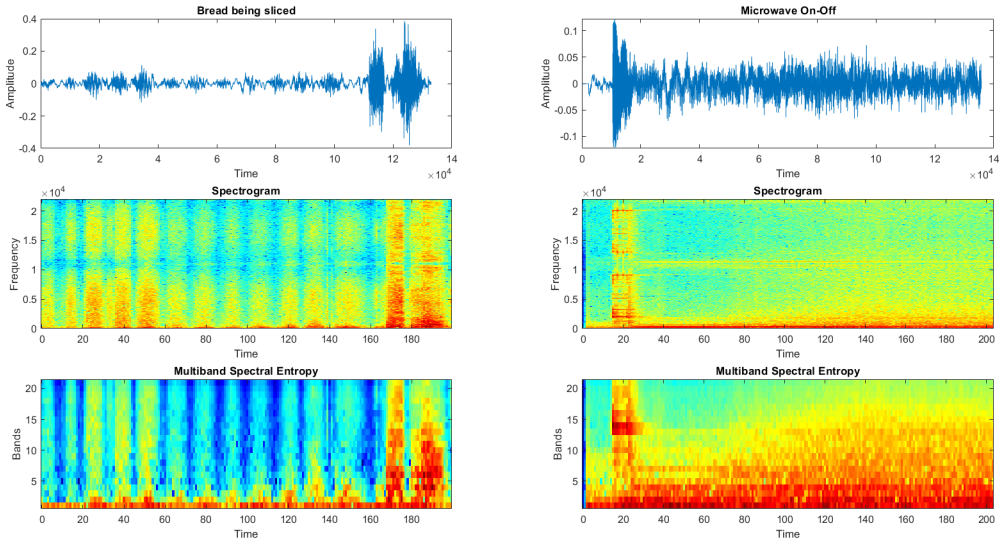

Figure 1: Illustration of MSES signatures with their corresponding signals and spectrograms. These signatures correspond to three seconds of audio from the acoustic events called "Bread being sliced" (left) and "Microwave On-Off" (right).

3. Database

An environment at home where different sound sources occur at the same time could be the kitchen. 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 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 [3]. 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 archives of audio have WAV format, 44100Hz of sampling frequency and coded to 16 bits. No copyright infringement was intended. The downloaded sounds are given in Table 2.

In an approach of audio signatures, it is common to use signatures of short duration, usually between one y fifteen seconds. All downloaded archives of audio have length of three seconds (we consider that three seconds of audio is enough to identify a sound from the environment). For building the database all the sixteen original sounds were mixed. 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 originals 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 originals sounds are considered the "signal" (the acoustic events to identify) and the elements of the dataset as the "noise". Figure 2 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, SVM and ANN. For the experiments with 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

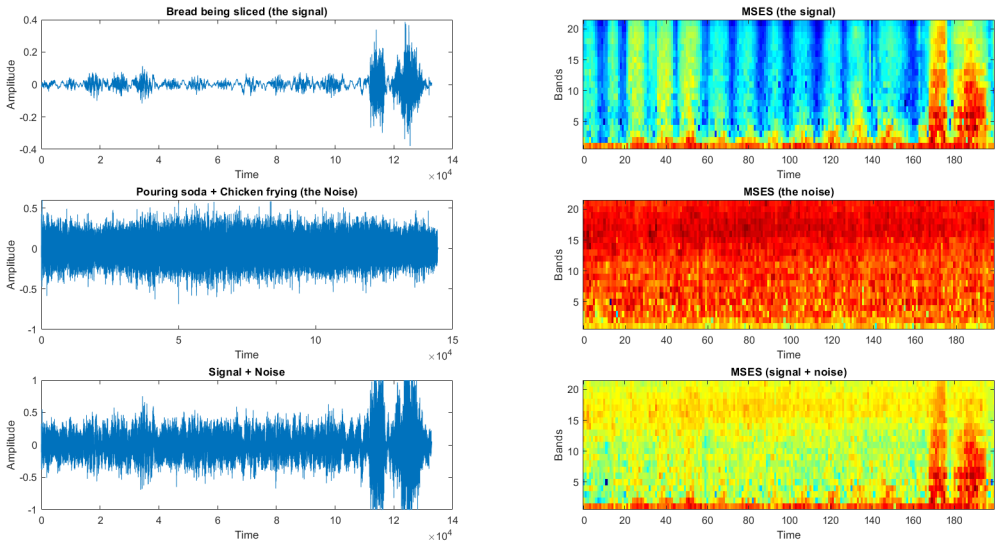

Figure 2: Illustration of the mixing process of sounds considering the acoustic event named "Bread being sliced" as the signal and the couple of sounds "Pouring Soda - Chicken Frying" as the noise. Left side shows the signals in time domain and right side shows the MSES signatures associated to every signal.

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`

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

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. All audios are cut to have 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) Then, a Hann window function is applied to each frame. e) By last, the FFT is computed for each frame.

From the previous, the procedure ends implementing the described in 2.1 and 2.3 Sections. 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

Our 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 each row of the $L \times T$ matrix by computing the mean and standard deviation, namely, $v_{ij} = (v^*_{ij} - \overline{\mu_i})/\sigma_i$, for $i = 1, 2, ..., L$ and $j = 1, 2, ..., T$, where v_{ij} denotes the i, jth normalized value, v^*_{ij} denotes the i, jth raw value of the signature, $\overline{\mu_i}$ denotes the mean of the ith row and σ_i denotes the standard deviation of the ith row.

Haitsma's work presents a method to binarize audio signatures. This method consists of taking the sign of the differences between consecutive values [12]. 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.

4.3. Experiment with Artificial Neural Networks

This experiment consists of training neural networks to classify the acoustic events that are considered the signal and not as 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 referred to MFCC and 48 signatures for training the neural network referred to MSES. Neural networks consist of 2 hidden layers and 16 neurons in the output layer; the input layer have 4200 neurons (i.e., each signature of size 21×200 is converted to vector). In addition, 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. For the first and second hidden layers, the hyperbolic tangent sigmoid transfer function is applied. For the output layer, the logarithmic-sigmoid transfer function is implemented. Finally, 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

Experiments with SVM use the same training dataset as with ANN. 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, Euclidean, Cityblock, 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 confusion matrices, the best experimental outcomes and the averages achieved with each classifier are summarized in Table 4.

5.1. Similarity Distance Results

For the results the recall metric was used. Table 2 shows the results for each signature using Hamming distance and Cosine distance. Although it is common to use binary signatures in an audio signature-based approach, the results of Table 2 allow to observe that binary signatures are not convenient to represent acoustic events, especially, when they have non-stationary characteristics. Summarizing, 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 works better, allowing to differentiate more the performance of both feature extraction methods, especially, when working with low levels of SNR.

For Hamming distance, results in Table 2 showed that C2 was the worst classified class getting zero in recall score, while, C3, C4, C13 and C16 are the classes of sounds that obtained the higher recall in both features, 100% in all of them. For Hamming distance, the average recall obtained by using MFCC features is 65.29% and 68.75% by using MSES features.

On the other hand, results with Cosine distance using MSES feature, it can be showed that C4, C8, C15 and C16 were the classes of sounds getting the higher recall score. Anew C2 was the worst classified class for both features. For Cosine distance, the average recall obtained by using MFCC

Table 2: Results about Recall

Acoustic Events[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.

features is 63.45% and 77.97% by using MSES features (i.e., the difference of recall between both features is about the 14.52%). The results of this experiment mark the baseline to know how much the scores improve by using machine learning methods.

One thing to note is that MSES captures the location of energy peaks in each sub-band that are less corrupted by noise, allowing them to work better than MFCCs for low SNR levels. However, both characteristics represent very well the non-stationary characteristics of audio signals. Therefore, the disadvantage for MFCC seems to be when working with audio signals that have a considerable degree of noise.

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 3: Results about 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

Figures 3 and 4 show 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 respectively. At least 14 samples of each class of kitchen sounds (except by C6) are classified erroneously as C5.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
C1	91	0	0	0	11	0	0	0	0	0	0	0	0	0	3	0
C2	3	37	2	0	14	5	12	6	0	0	0	4	6	1	10	5
C3	0	0	97	0	8	0	0	0	0	0	0	0	0	0	0	0
C4	5	0	0	56	13	6	0	8	0	4	0	0	5	7	0	1
C5	0	0	0	0	104	1	0	0	0	0	0	0	0	0	0	0
C6	2	0	0	0	0	103	0	0	0	0	0	0	0	0	0	0
C7	3	0	0	0	14	7	72	0	0	3	0	0	0	1	4	1
C8	0	0	0	0	14	2	0	88	0	1	0	0	0	0	0	0
C9	5	2	0	5	14	4	0	3	38	4	4	1	7	6	9	3
C10	3	0	0	0	12	1	0	0	0	89	0	0	0	0	0	0
C11	7	3	1	0	14	4	7	1	0	1	46	3	4	1	10	3
C12	0	0	0	0	14	2	0	0	0	0	0	89	0	0	0	0
C13	0	0	0	0	8	1	0	0	0	0	0	0	96	0	0	0
C14	4	0	0	0	14	3	0	0	0	2	0	1	4	75	1	1
C15	0	0	0	0	14	1	0	0	0	0	0	0	0	0	90	0
C16	0	0	0	0	12	4	0	0	0	0	0	0	0	0	0	89

Figure 3: Confusion Matrix with ANN model and MFCC features.

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

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
C1	98	0	0	0	0	0	0	0	0	0	0	0	0	6	1	0
C2	0	105	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C3	0	0	104	0	0	0	0	1	0	0	0	0	0	0	0	0
C4	0	0	0	105	0	0	0	0	0	0	0	0	0	0	0	0
C5	0	0	0	0	105	0	0	0	0	0	0	0	0	0	0	0
C6	0	3	0	0	0	81	0	0	0	0	1	0	4	3	12	1
C7	3	12	0	0	0	0	63	7	1	0	0	0	0	11	6	2
C8	0	0	0	0	0	0	0	104	0	0	0	0	0	1	0	0
C9	0	1	0	6	0	0	0	2	77	0	0	0	3	10	6	0
C10	0	0	0	6	0	0	0	0	0	90	0	0	0	8	0	1
C11	2	2	0	5	1	0	0	2	0	0	75	0	2	9	6	1
C12	0	0	0	0	0	1	0	2	0	0	0	101	0	0	0	1
C13	0	0	0	0	0	0	0	0	0	0	0	0	105	0	0	0
C14	0	0	0	0	0	0	0	0	0	0	0	0	0	105	0	0
C15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	105	0
C16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	105

Figure 4: Confusion Matrix with ANN model and MSES features.

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

Figures 5 and 6 show 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 with the MSES features 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

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
C1	68	0	0	0	14	0	0	3	0	0	0	3	5	0	11	1
C2	4	32	3	13	14	0	0	7	0	0	0	6	10	0	11	5
C3	0	0	93	0	12	0	0	0	0	0	0	0	0	0	0	0
C4	1	0	0	81	14	0	0	0	0	0	0	1	7	0	1	0
C5	0	0	0	0	105	0	0	0	0	0	0	0	0	0	0	0
C6	0	0	0	0	0	104	0	0	0	0	0	0	0	0	1	0
C7	6	2	0	13	14	1	28	4	1	3	1	4	11	0	11	6
C8	0	0	0	0	14	0	0	83	0	0	0	1	7	0	0	0
C9	1	0	0	11	14	0	0	5	43	2	0	4	10	0	12	3
C10	0	0	0	0	14	0	0	0	0	88	0	1	0	0	2	0
C11	5	1	0	13	14	0	0	2	0	1	40	6	9	0	11	3
C12	0	0	0	0	14	0	0	0	0	0	0	88	3	0	0	0
C13	0	0	0	0	11	0	0	0	0	0	0	0	94	0	0	0
C14	7	10	0	13	14	2	10	8	3	3	2	5	10	0	12	6
C15	0	0	0	0	14	0	0	0	0	0	0	0	0	0	91	0
C16	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	91

Figure 5: Confusion Matrix with SVM model and MFCC features.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
C1	103	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
C2	0	88	0	0	0	0	0	0	0	1	0	1	7	5	0	3
C3	1	0	104	0	0	0	0	0	0	0	0	0	0	0	0	0
C4	1	0	0	97	0	0	1	0	0	0	1	1	0	1	0	3
C5	7	0	0	0	59	0	14	0	0	0	7	0	5	11	0	2
C6	7	0	0	0	0	44	14	0	4	2	9	1	8	13	0	3
C7	3	0	0	0	0	0	90	0	0	0	0	1	0	10	0	1
C8	5	11	0	8	0	2	10	16	5	5	10	1	7	13	9	3
C9	0	0	0	2	0	0	0	0	89	1	0	1	4	8	0	0
C10	0	0	0	0	0	0	0	0	0	105	0	0	0	0	0	0
C11	2	1	0	0	0	0	0	0	0	0	93	1	0	8	0	0
C12	0	0	0	0	0	0	0	0	0	0	0	105	0	0	0	0
C13	0	0	0	0	0	0	0	0	0	0	0	0	105	0	0	0
C14	0	0	0	0	0	0	0	0	0	0	0	0	0	105	0	0
C15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	103	1
C16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	105

Figure 6: Confusion Matrix with SVM model and MSES features.

best performance with MFCC features was the model that uses the spearman distance function with two neighbors. Figure 7 shows the confusion matrix of this implementation. The recall metric was 65.77%. C3, C4, C5, C6 and C16 obtained the best results. Contrary, only one sample of C2 was correctly classified. The results obtained with MSES feature (Figure 8) 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.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
C1	58	0	0	12	12	2	0	0	0	0	0	0	9	0	10	2
C2	4	1	1	14	13	8	12	7	2	3	4	0	10	9	11	6
C3	0	0	104	0	0	0	0	1	0	0	0	0	0	0	0	0
C4	0	0	0	105	0	0	0	0	0	0	0	0	0	0	0	0
C5	0	0	0	0	104	1	0	0	0	0	0	0	0	0	0	0
C6	0	0	0	0	0	105	0	0	0	0	0	0	0	0	0	0
C7	2	0	0	13	13	3	60	0	0	0	0	0	9	0	4	1
C8	2	0	0	13	12	3	11	45	0	0	0	0	8	0	10	1
C9	0	0	0	13	13	5	10	1	39	2	0	0	11	0	9	2
C10	0	0	0	13	11	3	0	1	0	72	0	0	2	0	3	0
C11	3	0	0	13	13	6	11	4	1	11	19	1	10	0	10	3
C12	0	0	0	10	13	1	13	3	0	0	0	50	6	0	8	1
C13	0	0	0	0	0	0	0	0	0	0	0	0	105	0	0	0
C14	0	0	0	12	12	4	0	0	0	5	2	0	10	54	5	1
C15	0	0	0	0	10	0	0	0	0	0	0	1	0	0	94	0
C16	0	0	0	0	12	1	0	0	0	0	0	0	2	0	0	90

Figure 7: Confusion Matrix with KNN model and MFCC features.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
C1	104	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
C2	1	90	0	0	0	0	1	6	0	0	0	0	0	1	3	3
C3	0	0	103	0	0	0	0	1	0	0	0	0	0	0	0	1
C4	0	0	0	103	0	0	0	0	0	0	0	0	0	0	0	2
C5	7	0	0	0	73	0	14	5	0	0	0	1	0	0	3	2
C6	5	12	0	11	1	17	13	8	3	0	5	2	6	10	9	3
C7	1	0	0	0	0	0	100	3	0	0	0	0	0	0	0	1
C8	0	0	0	0	0	0	0	105	0	0	0	0	0	0	0	0
C9	0	0	0	7	0	0	0	2	85	0	0	0	3	3	4	1
C10	0	0	0	7	0	0	0	0	0	95	0	0	1	0	1	1
C11	4	0	0	0	0	0	1	7	0	0	72	1	1	8	9	2
C12	0	0	0	0	0	0	0	2	0	0	0	102	0	0	0	1
C13	0	0	0	0	0	0	0	0	0	0	0	0	105	0	0	0
C14	0	0	0	0	0	0	0	0	0	0	0	0	0	105	0	0
C15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	104	1
C16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	105

Figure 8: Confusion Matrix with KNN model and MSES features.

Table 4 shows the results obtained using MFCC and MSES features for Similarity distances, ANN, SVM and KNN. As expected, we can observe in all of them an improvement in the recall metric when working with MSES features. For both MFCC and MSES, the method of ANN was the one with the highest performance in our experiments (75% and 90.95%, respectively). Second best performance for MSES was achieved with KNN (i.e., 87.38%), this is because the nearest neighbors vote mostly for the sound that is the acoustic event to be identified. In third place the SVM method achieved

a recall of 83.99% and finally, similarity distances with a score of 77.97%. Regarding MFCC, the second best performance was achieved with SVM (i.e., 67.2%). 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 they work with variations that allow their learning to be more robust and effective than the other methods, in addition, a search was made for the best performance of the neural network in the learning stage by increasing one neuron from 10 up to 100 in the hidden layers.

Table 4: Best results for the classification tests 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

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 $\alpha = 0.05$. 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.

6. Conclusions

The goal of this work was to identify acoustic events using the approach of audio signatures. 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. Due to the above, deep learning techniques are beyond the scope of this work, since these methodologies are usually used when there are several instances of the sound classes. Two audio features were considered in this work, MFCC and MSES, where the first is our benchmark feature and the second our proposal. We use MFCC as reference because is the most cited audio feature when working

with audio-based activity recognition. The results showed that the representation of acoustic events based on MSES is more convenient when working with different classification methods. Although the competition between MSES and MFCC is not absolute, it seems that MSES is an audio feature that is very robust for identifying acoustic events in a mixture of sounds. 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 the databases found in the literature are different in the sense that they do not consider the mixing of sounds from the same scenario, in instead, they are constructed with sounds from different scenarios (for instance, task 4 refers to speech, dog, dishes,blender, electric shaver, etc. without mix them) such as the DCASE2019 database in its different tasks. The goal of the experiments was to evaluate MSES and MFCC looking for those acoustic events that are found in the mixtures of sounds contained in the database. In real home environments different activities generate different types of sounds that can be mixed, therefore, it is important to investigate new methodologies that automatically identify all the sounds that are present in the environment in order to give assistance to people, among other tasks. 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 it is important for recognize activities in real indoor environments.

Acknowledgments

This work was supported by PRODEP and UACH – FING through the project provided to the first author with title “Diseño de un Sistema Embebido de Monitoreo de Actividad y de Métricas de Estado Vascular para Vida Asistida”.

References

[1] 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. Journal of the Acoustical Society of America .

[2] Battaglino, D., Mesaros, A., Lepauloux, L., Pilati, L., Evans, N., 2015. Acoustic context recognition for mobile devices using a reduced complexity svm. 23rd European Signal Processing Conference , 534–538.

[3] Beltrán-Márquez, J., Chávez, E., Favela, J., 2012. Environmental sound recognition by measuring significant changes in the spectral entropy, in: A., C.O.J., F., M.T.J., A., O.L.J., K.L., B. (Eds.), Pattern Recognition. MCPR 2012. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.

[4] Camarena-Ibarrola, A., Chávez, E., 2006. On musical performances identification, entropy and string matching. Mexican International Conference on Artificial Intelligence , 952–962.

[5] Camarena-Ibarrola, A., Chávez, E., Sadit-Tellez, E., 2009. Robust radio broadcast monitoring using a multi-band spectral entropy signature. Iberoamerican Congress on Pattern Recognition , 587–594.

[6] Camarena-Ibarrola, A., Figueroa, K., Tejeda-Villela, H., 2016. Entropy per chroma for cover song identification. IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC) .

[7] Camarena-Ibarrola, A., Luque, F., Chávez, E., 2017. Speaker identification through spectral entropy analysis. IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC) .

[8] Chachada, S., Kuo, C.C.J., 2013. Environmental sound recognition: A survey. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference .

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

[10] Cowling, M., Sitte, R., 2003. Comparison of techniques for environmental sound recognition. Pattern Recognition Letters 24, 2895–2907.

[11] Gencoglu, O., Virtanen, T., Huttunen, H., 2014. Recognition of acoustic events using deep neural networks. 22nd European Signal Processing Conference (EUSIPCO) , 506–510.

[12] Haitsma, J., Kalker, A., 2002. A highly robust audio fingerprinting system. Proceedings of International Symposium on Music Information Retrieval .

[13] Janvier, M., Alameda-Pineda, X., Girin, L., Horaud, R., 2012. Sound-event recognition with a companion humanoid. 12th IEEE-RAS International Conference on Humanoid Robots .

[14] Jia-Lin, S., Jeih-Weih, H., Lin-Shan, L., 1998. Robust entropy-based endpoint detection for speech recognition in noisy environments. 5th International Conference on Spoken Language Processing 98, 232–235.

[15] Khunarsal, P., Lursinsap, C., Raicharoen, T., 2013. Very short time environmental sound classification based on spectrogram pattern matching. Information Sciences 243, 57–74.

[16] Kücükbay, S.E., Sert, M., 2015. Audio-based event detection in office live environments using optimized mfcc-svm approach. IEEE 9th International Conference on Semantic Computing , 475–480.

[17] Logan, B., 2000. Mel frequency cepstral coefficients for music modeling. IEEE Proceedings of the International Symposium on Music Information Retrieval .

[18] Manzo-Martínez, A., Camarena-Ibarrola, A., 2015. Use of the entropy of a random process in audio matching tasks. 38th International Conference on Telecommunications and Signal Processing (TSP) .

[19] Martín-Morató, I., Cobos, M., Ferri, F.J., 2016. A case study on feature sensitivity for audio event classification using support vector machines. IEEE International Workshop on Machine Learning for Signal Processing .

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

[21] Misra, H., Ikbal, S., Bourland, H., Hermansky, H., 2004. Spectral entropy based feature for robust asr. Proceedings of International Conference on Acoustics, Speech and Signal Processing , 193–196.

[22] Misra, H., Ikbal, S., Sivadas, S., Bourlard, H., 2005. Multi-resolution spectral entropy feature for robust asr. Proceedings of International Conference on Acoustics, Speech, and Signal Processing , 253–256.

[23] Mohammad, A., 1994. Entropy in signal processing. Traitement du Signal , 87–116.

[24] Mohanapriya, S.P., Sumesh, E.P., Karthika, R., 2014. Environmental sound recognition using gaussian mixture model and neural network classifier. International Conference on Green Computing Communication and Electrical Engineering .

[25] Sang-Wook, P., Jin-Sang, R., Min-Kyu, S., Han, D., Hanseok, K., 2014. Acoustic feature extraction for robust event recognition on cleaning robot platform. IEEE International Conference on Consumer Electronics .

[26] Schroeder, J., Wabnik, S., Van Hengel, P.W.J., Goetze, S., 2011. Detection and classification of acoustic events for in-home care, in: R., J., B., E. (Eds.), Ambient Assisted Living. Springer, Berlin, Heidelberg, pp. 181–195.

[27] Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27, 379–423.

[28] Sigurdsson, S., Petersen, K.B., Lehn-Schiøler, T., 2006. Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. 7th International Conference on Music Information Retrieval .

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

[30] Stork, J.A., Spinello, L., Silva, J., Arras, K.O., 2012. Audio-based human activity recognition using non-markovian ensemble voting. The 21st IEEE International Symposium on Robot and Human Interactive Communication , 509–514.

[31] Takahashi, N., Gygli, M., Pfister, B., Gool, L., 2016. Deep convolutional neural networks and data augmentation for acoustic event detection. Interspeech .

[32] Temko, A., Nadeu, C., 2006. Classification of acoustic events using svm-based clustering schemes. Journal Pattern Recognition 39(4), 682–694.

[33] Traunmüuller, H., 1990. Analytical expressions for the tonotopic sensory scale. The Journal of the Acoustical Society of America 88, 97–100.

[34] Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., Sarti, A., 2007. Scream and gunshot detection and localization for audio-surveillance systems. Conference on Advanced Video and Signal Based Surveillance .

[35] Zhang, H., McLoughlin, I., Song, Y., 2015. Robust sound event recognition using convolutional neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing .