Novel Time-Frequency Based Scheme for Detecting Sound Events from Sound Background in Audio Segments

Vahid Hajihashemi[1(✉)], Abdorreza Alavigharahbagh[1], Hugo S. Oliveira[1], Pedro Miguel Cruz[2], and João Manuel R. S. Tavares[3]

[1] Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
Hajihashemi.vahid@ieee.org
[2] Bosch Security Systems S.A., Ovar, Portugal
[3] Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

Abstract. Usually, Sound event detection systems that classify different events from sound data have two main blocks. In the first block, sound events are separated from sound background and in next block, different events are classified. In recent years, this research area has become increasingly popular in a wide range of applications, such as in surveillance and city patterns learning and recognition, mainly when combined with imaging sensors. However, it still poses challenging problems due to existent noise, complexity of the events, poor microphone(s) quality, bad microphone location(s), or events occurring simultaneously. This research aimed to compare accurate signal processing and classification methods to suggest a novel method for detecting sound events from sound background in urban scenes. Using wavelet and Mel-frequency cepstral coefficients, the analysis of the effect of classification methods and minimization of the number of train data are some of the advantages of the proposed method. The proposed methods' application to a standard sounds database led to an accuracy of about 99% in event detection.

Keywords: Signal processing · Wavelet transform · Machine learning · Event detection

1 Introduction

Information processing algorithms are a paramount step in artificial intelligence growth. However, the current modes of human-machine communication are geared more towards living with the limitations of computer input/output devices, mainly as to sound and image, rather than the convenience of humans. Sound is one of the primary modes of communication among humans or between environment and humans. On the other hand, it would be interesting if computers could listen to sound and understand meanings. Automatic speech recognition and sound event detection are processes of deriving the word sequence or sound reason, given the speech waveform. Speech understanding goes one step

J. M. R. S. Tavares et al. (Eds.): CIARP 2021, LNCS 12702, pp. 402–416, 2021.
https://doi.org/10.1007/978-3-030-93420-0_38

further, and describe the meaning of the signal in terms of human communication. Intelligent agents such as mobile phones, hearing aids or robots could also hear, but they cannot exactly interpret what is heard. Sound is often a supplement to content such as video and contains information about the environment. The difference is that often the sound can be collected and processed in easier ways. The information gathered from a meaningful sound analysis can be useful for other processes such as robot routing, user alerting, or analysis and understanding details of an event.

A sound event is a designation commonly used to describe a recognizable event in an audio segment. This designation usually enables a person to understand the meaning of an event and how it relates to other events. Sound events can be used to represent a scene symbolically; for example, a hearing scene on a busy street includes cars passing, cars crash and footsteps of pedestrians. Sound scenes can be described with different specific sound events to assign the main subject, for instance, a street Semantic and automatic event detection and understanding are fundamental requirements in modern urban surveillance systems towards smarter and safer cities. While such systems rely heavily on imaging data, other types of data, such as audio data, can be used to overcome the weaknesses of the visual-based systems and enhance the outcomes of the systems towards better decision making by the city authorities. Because of the fuzzy nature of sound events interpretation, artificial intelligence still has many weaknesses in comparison to the human system. Based on challenges in sound event detection in urban scenes, in this article, a comprehensive analysis of usual state of the art features in sound event detection is presented, and a novel time-frequency method is suggested to automatically detect sound events from sound background in audio segments acquired in urban scenes. This article is structured as follows: the next section gives an overview of state-of-the-art researches in sound event detection. The third section describes the mathematical and theoretical fundamentals of wavelet transform (WT), Mel-frequency cepstral coefficients (MFCCs), K-nearest neighbor (KNN) and support vector machine (SVM), which are used in the proposed method. Section 4 presents details of the used database and then describes the proposed method. Simulation results and conclusions are given in Sects. 5 and 6, respectively.

2 Literature Review

In recent years, efforts have been made to expand the issue of sound event recognition to a comprehensive set of events in environments. Most audible scenes are complex in terms of events, as they usually involve several simultaneously active overlapping sound events. There are two ways to automatically detect a sound event: 1) Finding the start and end time of an event in an audio segment, and then make a single-channel (Monophonic) sequence including events as output [1]. This method is called single-channel detection (Monophonic Detection). 2) Finding some events in a multi-channel (Polyphonic) sequence of events, which is called multi-channel recognition [2,3]. A lot of research has been done in sound

event detection. An unrelated field approach was proposed in [4], which consists of two steps: automatic background detection and sound event detection. A method for modeling the previous probabilities of overlapping events was proposed using a probabilistic latent semantic analysis (PLSA) to calculate previous probabilities and learn the relationships between event sources [2].

There are two ways to detect events in multi-source environments that can detect multiple overlapping sound events. The first uses uniform single-channel recorded sounds (mixed signals), and in the detection stage, uses several limited Viterbi [5] transitions to record overlapping events [6]. The second uses Unsupervised Source Selection as a processing step to minimize the impact of overlapping events, and the detection step is performed separately for each system, [7]. In [3], two methods based on an iterative algorithm for Exception Maximization (EM) used to select the desired voice: one, based on the most probable current selection; and another, based on the gradual elimination of the most probable current from the training. The relationship between sound and label in a sound database was studied by evaluating the semantic similarity of sample labels with similar semantic sounds in [8]. On the other hand, a method for combining sound similarity and semantic similarity in a single similarity criterion was proposed in [9].

In some research, the audio signal is recognized as a single-channel signal with one event at a time [10,11]. LeCun et al. proposed a system for detecting an event in a real-life recorded file, using deep learning [12]. In [2], a Probabilistic Latent Semantic Analysis (PLSA), a method close to Non-negative Matrix Factorization (NMF), was proposed to detect overlapping sound events. Simultaneously with the occurrence of events, the degree of overlap of a polyphonic part is represented. Cotton and Ellis applied NMF to MFCCs and tested proposed method on the detection of heterogeneous sound events [10]. In speech recognition applications, a usual assumption is the existence of a dominant source that should be analyzed [13], but this assumption is not true in event detection. One strategy to manage multi-voiced signals is to separate sound resources and analyze each source separately [7,14]. In [14], a study on computational analysis of auditory scenes was performed to study human-robot interaction by recognizing auditory information. A review of the latest research in the category of sound event categorization is presented in [15], where various types of convolutional neural network (CNN) architectures used to categorize sound events are described.

Many researchers have focused on sound denoising to increase the sound event detection accuracy. According to our review, considerable research has been done on urban noise modeling and not so on noise removal. Usual noise management approaches are focused on the reduction of the noise energy [16]. In event detection, noise is usually defined as any unwanted normal environment sound that may decrease the accuracy of the abnormal sound detection. Two categories have been introduced for noise removing: energetic masking (EM) and informational masking (IM). EM uses similar time-frequency locations [17,18] and has weakness in high-energy [17,19]; therefore, EM alone is not a good choice [20]. IM is an indirect saliency-based method [21,22] of auditory attention [23,24].

A wide range of features in different domains has been used to detect sound events, namely: Spectrogram [25], patterns similarity in the time domain [26,27], and spectrum as suggested in [28]. Linear predictive coding was used in [29] for sound-based rare-event detection. Mel scale [30], Discrete Cosine Transform [31,32], Mel-Frequency Cepstral Coefficients [33,34], Wavelet decompression [35], Perceptual linear prediction (PLP) [36], Linear prediction cepstral coefficients (LPCCs) [37], and Line spectral frequencies (LSFs) [38] are other features that have been used for sound event detection in different researches.

Each of the aforementioned features has its weaknesses and advantages and none of the mentioned research works were able to specify a feature as the best in sound event detection. Based on our review, some challenges in sound event detection are the presence of different events in one soundtrack, unbalanced number of event data versus normal data in the training process, dynamic context-dependent form and different speed of occurrence. Recent researches usually use MFCCs and Wavelet based features as better features. Various classification methods can be used (based on differences between features) for sound event detection and understanding. Some of the most well-recognized classification methods that have been used in this topic are Logistic Regression, SVM, KNN, Fuzzy C-means clustering, Adaptive Neuro-Fuzzy Inference Systems, Naïve-Bayes, and Deep learning (mainly, Convolutional Neural Networks). In the proposed method, MFCCs and wavelet were selected as feature extractors, and a novel statistical scheme based on normalized histogram is used for feature processing. In the second step, SVM and KNN are used as detection methods.

3 Theoretical Framework

3.1 Wavelet Transform

One of the common feature extraction methods in sound processing is WT. In practice, audio signals are time-domain signals in their raw format. That is, whatever the signal is conveying, is a function of time. In many cases, the most distinguished information is hidden in the frequency content of the signal. The need for WT arises because in sound events, is necessary to have both the time and the frequency information at the same time depending on the particular application, and the nature of the signal in hand, since no frequency information is available in the time-domain signal, and no time information is available in the Fourier space [2,3].

The basic element of WT is known as the "mother wavelet" function, $\Psi(t)$. The Fourier transform of $\Psi(t)$, which is defined as $\Psi(\omega)$, must satisfy the following condition:

$$\int_{-\infty}^{+\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega = C_\Psi < +\infty. \tag{1}$$

Performing scaling and translation operations on $\Psi(t)$ creates a family of scaled and translated versions of the mother wavelet function:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right), \tag{2}$$

where a is the scaling and b is the translation parameters, respectively. Given a mother wavelet function $\Psi(t)$, the continuous wavelet transform (CWT) of function $f(t)$ is:

$$CWT_f\,(a,b) = |a|^{-1/2} \int_{-\infty}^{\infty} f\,(t)\,\psi^*\left(\frac{t-b}{a}\right) dt \;\; a,b \in R, a \neq 0, \tag{3}$$

where $*$ denotes the complex conjugate. For discrete wavelets, scale-time parameters a and b are discretized as $a = a_0^m$ and $b = nb_0a_0^m$.

This family of mother discretized wavelet functions $\{\Psi_{m,n}\,(t)\}$ is given as:

$$\Psi_{m,n}\,(t) = a_0^{-m/2}\Psi\left(a_0^{-m}t - nb_0\right) \qquad m,n \in Z. \tag{4}$$

So, by using Eq. (4), Eq. (3) can be rewritten as:

$$(DWT_f)_{mn} = a_0^{-m/2} \int_{-\infty}^{\infty} f\,(t)\,\psi\,\left(a_0^{-m}t - nb_0\right) dt. \tag{5}$$

The mother wavelet functions used in this work are:

$$\Psi(t) = \frac{2}{\sqrt{3}}\pi^{-\frac{1}{4}}\left(1-t^2\right)\exp\left(-\frac{t^2}{2}\right), \tag{6}$$

$$F_n\,(t) = \sum_{j=-n}^{n}\left(1 - \frac{|j|}{n+1}\right)e^{ijt} = \frac{1}{n+1}\left\{\frac{\sin\frac{n+1}{2}t}{\sin t/2}\right\}^2, \tag{7}$$

$$\Psi\,(t) = \begin{cases} 1 & 0 \leq t < 0.5 \\ -1 & 0.5 \leq t < 1 \\ 0 & otherwise \end{cases}. \tag{8}$$

3.2 Mel-Frequency Cepstral Coefficients

MFCCs are based on the known variation of the human ear's critical bandwidths. In MFCC, some filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of sound signals. The Mel-frequency scale is a combination of linear frequency spacing 1000 Hz and a logarithmic spacing 1000 Hz. A block diagram of the structure of a MFCC scheme is given in Fig. 1. The audio signal is typically recorded at a sampling rate above 10 kHz. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are heard by humans. The main purpose of the MFCCs is to mimic the behavior of the human ears. In addition, rather than the sound waveforms themselves, MFCC's are shown to be less susceptible to noise.

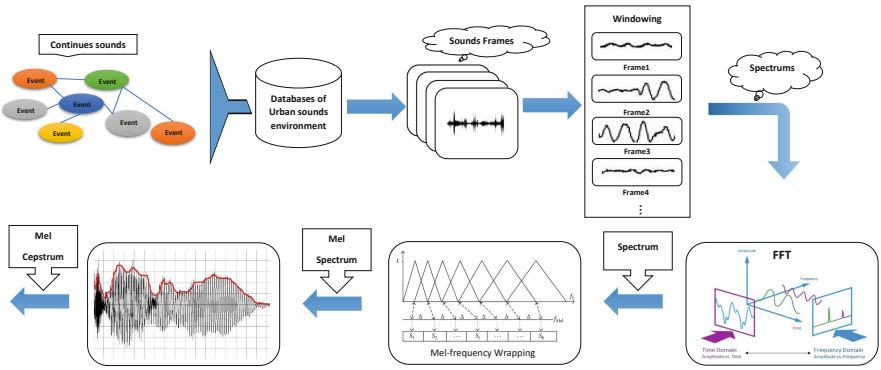


Fig. 1. Block diagram of the MFCC scheme.

3.3 Support Vector Machine

Since SVM classifiers are suitable for binary classification, in this study, a SVM is used for building a binary classifier between any sound event and sound background in audio segments. For the SVM classifier, different kernels were tested. SVM classifies the input data by building an imaginary hyperplane based on its kernel and tries to maximize the margin of that hyperplane to build a safe boundary for binary classification, as well as helping to find non-linear data pattern to classify input. Given a training set of N data points $\{y_k, x_k\}_{k=1}^{N}$ where $x_k \in R^n$ is the k-th input pattern and $y_k \in R$ is the k-th output pattern, the classifier can be constructed using the SVM method in the form:

$$y(x) = sign\left[\sum_{k=1}^{N} \alpha_k y_k K(x, x_k) + b\right], \tag{9}$$

where α_k is a non-negative Lagrange multiplier, b is a constant, and $K(\cdot,\cdot)$ is the kernel, which can be either $K(x, x_k) = x_k^T x$ - linear SVM, $K(x, x_k) = (x_k^T x + 1)^d$ - polynomial SVM of degree d, $K(x, x_k) = \tanh[\kappa\, x_k^T x + \theta\,]$ - multilayer perceptron SVM, or $K(x, x_k) = \exp\{-\|x - x_k\|_2^2 / \sigma^2\}$ - RBF SVM, where κ, θ and σ are constants. First, a safety margin (Λ) is defined as:

$$\begin{aligned} if(x \in class \quad 1) &\Rightarrow \sum_{k=1}^{N} \alpha_k y_k K(x, x_k) + b \geq \Lambda, \\ if(x \in class \quad -1) &\Rightarrow \sum_{k=1}^{N} \alpha_k y_k K(x, x_k) + b \leq -\Lambda. \end{aligned} \tag{10}$$

The SVM training step uses kernel function parameters, α_ks and b, to maximize Λ and the total accuracy. Many analytical, numerical and heuristic methods have been suggested for finding α_ks, b and the selected kernel parameters using training data.

3.4 K-Nearest Neighbor Classifier

KNN is a simple and non-parametric supervised classification algorithm that can be useful for classification and regression problems. In KNN classification, the output can be defined as a multiclass output. An object is classified by computing its distance to some known centers and the object is assigned to the class most similar, near or common among its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the closest nearest center. In KNN regression, the output is the property value for the object, which is the mean of the values of k nearest neighbors. The number of neighbors and similarity or distance metric are the main factors of KNN. The distance measure can be selected as Euclidean, Hamming, Manhattan, or Minkowski distance. In this research, the centers for each event were separately found using k-means clustering. The distance of new input were compared to all centers and each one assigned to background or event.

4 Proposed Method

The train step pseudo-code of the proposed system is given by Algorithm 1.

Algorithm 1: *Training procedure*

Input: Labelled recorded signal from urban scenes **(S_i)**
  Split recordings into one-second non-overlapping sections **(S_{si})**
    Apply one-dimensional wavelet transform to **S_{si}** and make two output signals **(cA_{ssi} , cD_{ssi})**
    Reshape the outputs to N × 8 matrices
    Calculate the 16-bin normalized histogram of two output signals (According to columns) and make 16×16 feature matrix
    Assign event and non-event label to each feature matrix
  Train classifier (**SVM**)
Output: Trained classifier

For sound event detection, many databases were collected and labelled by humans. In the evaluation of our method, the US-SED dataset [39,40] was used because, at first, it is easily accessed and converted to different software formats and, second, it covers several important urban sound events.

4.1 The US-SED Dataset

US-SED is a large dataset of 10,000 ten-second soundscapes and includes ten different sound classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren and street music, which has been used for training and evaluating Sound Event Detection (SED) algorithms. All soundscapes were extracted from the UrbanSound8K dataset, approximately 1000 per each of ten urban sound sources (each clip contains one of the ten sources), as the soundbank. UrbanSound8K is pre-sorted into 10 stratified folds, and so can be used as folds 1–6 for generating 6000 training soundscapes, 7–8 for generating 2000 validation soundscapes, and 9–10 for generating 2000 test

soundscapes. Soundscapes were generated using the following strategy: first, a background sound normalized to -50 LUFS (Loudness Units relative to Full Scale) was added. The same background sound file for all soundscapes was used in combination with a 10-second clip of Brownian noise, which resembles the typical "hum" often heard in urban environments. By using a purely synthesized background, the database maker was guaranteed that it does not contain any spurious sound events that would not be included in the annotation. Next, the label was chosen randomly from all 10 available sound classes, and the source file was chosen randomly from all clips matching the selected label [39,40].

4.2 Pre-processing

In the pre-processing, the steps of removing noise and dividing the audio signal into non-overlapping segments are performed. In the first step, signals with a frequency of less 20 Hz and above 20 kHz, whose range is outside the human hearing range, are removed. Hence, all noises outside the hearing frequency band, which may have been transmitted to the signal, are eliminated. In the second step, the event situation is assumed constant in each non-overlapping segment. To make this assumption correct, the audio signals are divided into segments of typically 100 ms without overlap. The classifier then classifies the features extracted from these segments. In other words, the event in one tenth second signal is supposed constant. Based on this assumption, each incoming sound is broken into non-overlapping segments with one tenth second length. After splitting the sound into non-overlapping segments, each audio segment is labeled related to including or non including a sound event. In this case, a binary vector is made, which was established based on event segments versus background segments in the database. The 0 (zero) is equivalent to the background segment, and one shows that a sound event has occurred. After denoising, splitting and labeling steps, the audio signal is ready to enter the feature extraction block.

4.3 Feature Extraction

At this block, WT is applied to the input audio signal in order to decompose it and obtain the approximation and detail coefficients. A total of 15 mother wavelet functions with different parameters were implemented in the feature extraction block and studied in terms of the accuracy and efficiency. The used wavelet functions are indicated in Table 1.

Due to the sampling frequency of the input audio signal, which was 44100 Hz, enough samples are available for wavelet. The coefficients of two approximations coefficients cA1 and detail coefficients cD1 resulting from the WT are arranged into two $8 \times N$ matrices and then merged in a $16 \times N$ matrix. For each row, the probability density function (PDF) of the amplitudes is calculated. The number of intervals for PDF is 16 and the length of all intervals are supposed the same. Finally, the output of all rows is added together and arranged as a vector with length equal to 256. Obviously, the length of the feature vector is equal for all types of wavelet functions. The second approach is MFCC. In MFCC,

Table 1. Designation and number of used wavelet functions.

Wavelet designation	Wavelet number
Haar	–
Daubechies	10, 20
Symlet	2, 10, 20
Coiflet	1
Discrete Meyer	–
Fejer-Korovkin	4, 8, 22
Reverse biorthogonal	1.1, 2.4
Biorthogonal	1.1, 2.4

firstly the short-term Fourier transform (STFT) is applied to the signal using the Hanning window. The selected Hanning window is considered periodic and its length is equal to 512. In addition, 128 overlapping samples are considered for each segment. The output of the short-time Fourier transform is applied as input to the MFCC feature extraction step, and according to the sampling frequency available in the database, the feature-length of the MFCC output is a 13×11 matrix. Each row of this matrix is analyzed using a PDF, according to 20 intervals, in order to have the closest adaptation to the wavelet features. The final feature vector of MFCC has 260 elements. The feature vectors are input of detection block.

4.4 Detection

In this study, due to the very high imbalance problem between the background sound segments and the ones with specified sound events, a KNN or a binary SVM is trained to separate background from event sound segments. In this case, the classifier does not need to know type of events occurred in one tenth second segment; therefore, an audio segment is tested in such a way that it can include any event or has only background sound. Simulations were performed to evaluate the accuracy of the different SVM kernels and KNN on different feature types, including mother wavelet functions and MFCCs. After training, the trained classifier should be tested. The pseudo code of the test step is given by Algorithm 2.

Algorithm 2: *Test procedure*

Input: Labelled recorded signal from urban scenes ($\mathbf{S_i}$)
Split recordings into one-second non-overlapping segments ($\mathbf{S_{si}}$)
Apply one-dimensional wavelet transform to $\mathbf{S_{si}}$ and make two output signals ($\mathbf{cA_{ssi}}$, $\mathbf{cD_{ssi}}$)
Reshape the outputs to N × 8 matrix
Calculate the 16-bin normalized histogram of two output signals (according to columns) and make a 16×16 feature matrix
Apply feature matrix to the trained classifier
Calculate the method accuracy for test data
Output: Accuracy

5 Simulation Results

In order to determine whether WT can be a good option for separating events from background sound in audio segments, the time domain and two channels of corresponding WT for four different classifiers were considered, Fig. 2.

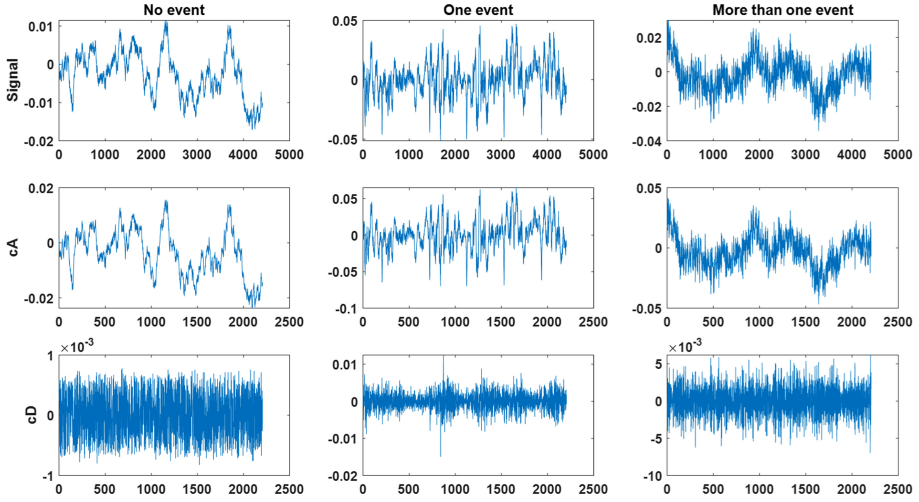
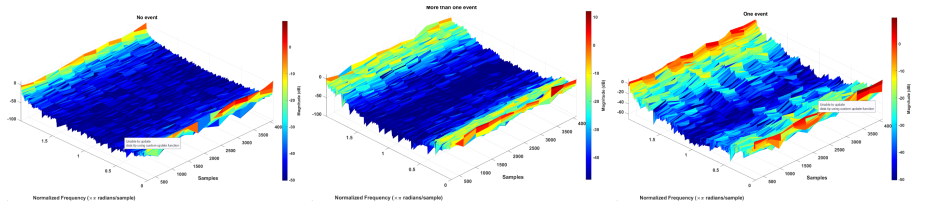


Fig. 2. Original audio signal and corresponding wavelet output for background and event sound segments.

As can be seen in Fig. 2, when the number of events increases, the signal pattern is compressed over time. Especially in cD, the amplitude and shape changes are obvious. These changes can be seen in time domain, but in WT, the changes are even clearer. It should be noted that all three audio segments in Fig. 2 were selected from the same environment and microphone, so the background noise and sound recording conditions are exactly similar for the three signals. To show the difference between event and background sound segments in MFCC,

the short-time Fourier transform of the signals were extracted using Hanning window as a main block of MFCC, Fig. 3.



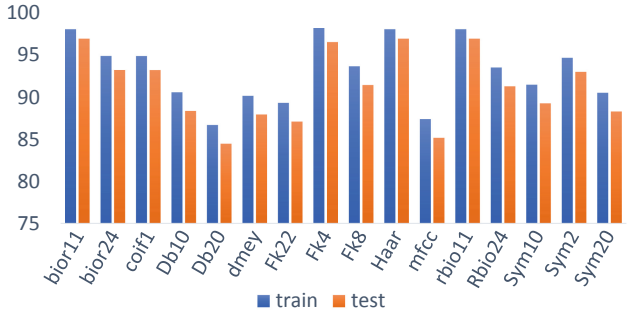
(a) (b) (c)

Fig. 3. (a) 3D STFT output for a background sound segment; (b) 3D STFT output for a segment including one sound event, and (c) 3D STFT output for a signal including more than one sound event.

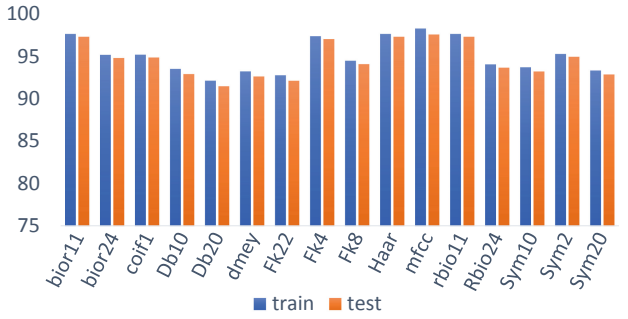
Given the three-dimensionality of the analysis, it can be seen that in a no event segment, the amplitude range is between −50 and 5 dB, and the maximum signal strength is located in a narrow band at the end of bands. After a sound event occurred in the environment, the energy is expanded to the intermediate bands, and the signal pattern has completely changed relative to the signal without sound events (Fig. 3). Additionally, It is more compact and the amplitude is changed significantly when more than one sound event occurred. It can be seen that in the field of Mel coefficients, the signal changes are quite obvious in the background and event sounds, which can be used as a criterion for detecting the background signal from including events signal. In the training classifier step, for eliminating the effect of a random selection of training data, the classifier was trained 20 different times using different wavelet functions, and the average accuracy of 20 times is reported as the final value. It should be noted that in each training, the train and test samples were similar for all feature types. The results obtained using KNN and different SVM kernels are given in Fig. 4. Due to a large number of samples, only 10% of the existing 1 million segments (100,000 samples) were used for training, and all the remaining data was used as test data. Although the percentage of training data is low, it can be seen that the accuracy was still very good, and the trained system was able to detect background vs event sounds in the audio segments.

In Fig. 4, the accuracy for KNN and studied SVM kernels, including linear, RBF and Polynomial, is given separately for train and test sets. In RBF and Polynomial cases, Mel's coefficients, which have been cited as best feature in various researches, showed better accuracy than other methods, although, in linear kernel and KNN, Mel coefficients did not work well due to their non-linear properties.

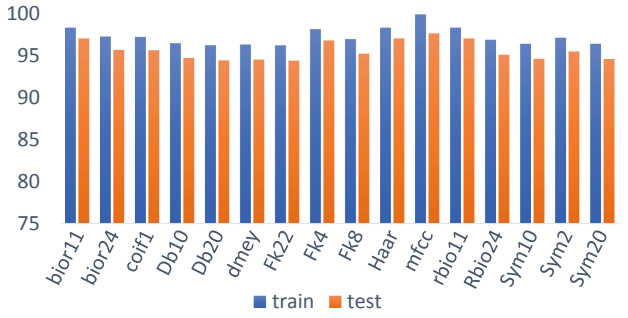
Among each of the studied kernels, the best accuracy was obtained using MFCC with Polynomial Kernel (99.9%), which indicates the very good accuracy in separating the event signal from the background signal. Figure 4, shows that



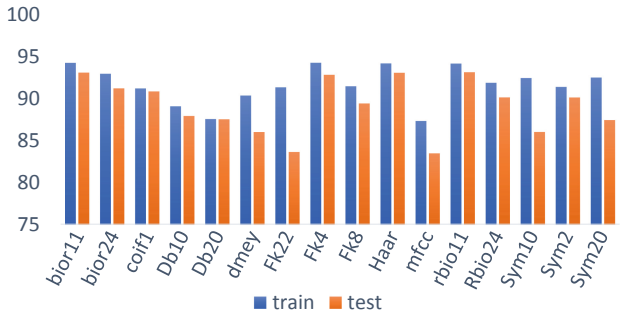
(a)



(b)



(c)



(d)

Fig. 4. Accuracy obtained by three different SVM kernels: a) linear, b) RBF, c) Polynomial, and d) KNN.

RBF and polynomial kernels had better accuracy than the linear kernel and KNN. In the meantime, RBF shows closer values in train and test sets (difference between test and train accuracy in all cases was lower than 1%), so for training stability, RBF is better than polynomial. In addition, the results indicate that the system can be easily trained using a small percentage of the available data.

6 Conclusion

In this study, a method based on efficient sound features and classifiers, which can be used in urban scenarios to differentiate the presence of sound events from the background sound in audio segments is proposed. The method uses Mel's coefficients and WT in combination with normalized histogram to separate sound events from the background sound with good accuracy. In the training step, SVM with Polynomial kernel showed the best accuracy, which was equal to 99.9%. As to the training stability, SVM with RBF kernel showed the closest values in train and test sets. Therefore, the proposed method can be used as a pre processing step to separate sound events from background sound in sound event classification systems. The simplicity and the good accuracy are among the advantages of the proposed method.

Acknowledgement. This article is a result of the project Safe Cities - "Inovação para Construir Cidades Seguras", with reference POCI-01-0247-FEDER-041435, co-funded by the European Regional Development Fund (ERDF), through the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), under the PORTUGAL 2020 Partnership Agreement.

References

1. Heittola, T., Mesaros, A., Eronen, A., Virtanen, T.: Context-dependent sound event detection. EURASIP J. Sound Speech Music Process. **1**, 1–13 (2013)
2. Mesaros, A., Heittola, T., Klapuri, A.: Latent semantic analysis in sound event detection. In: 2011 19th European Signal Processing Conference, pp. 1307–1311. IEEE (2011)
3. Heittola, T., Mesaros, A., Virtanen, T., Gabbouj, M.: Supervised model training for overlapping sound events based on unsupervised source separation. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8677–8681. IEEE (2013)
4. Heittola, T., Mesaros, A., Eronen, A., Virtanen, T.: Context-dependent sound event detection. EURASIP J. Sound Speech Music Process. **1**(1), 1 (2013)
5. Forney, G.D.: The viterbi algorithm. Proc. IEEE **61**, 268–278 (1973)
6. Eddy, S.R.: Hidden Markov models. Curr. Opin. Struct. Biol. **6**, 361–365 (1996)
7. Heittola, T., Mesaros, A., Virtanen, T., Eronen, A.: Sound event detection in multisource environments using source separation, in Machine Listening in Multisource Environments (2011)
8. Mesaros, T.H., Palomäki, K.J.: Analysis of acoustic-semantic relationship for diversely annotated real-world sound data. In: ICASSP, pp. 813–817 (2013)

9. Mesaros, T.H., Palomäki, K.: Query-by-example retrieval of sound events using an integrated similarity measure of content and label. In: 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pp. 1–4 (2013)
10. Cotton, V., Ellis, D.P.: Spectral vs. spectro-temporal features for acoustic event detection. In: Applications of Signal Processing to Sound and Acoustics (WASPAA), pp. 69–72 (2011)
11. Mesaros, A., Heittola, T., Virtanen, T.: Metrics for polyphonic sound event detection. Appl. Sci. **6**(6), 162 (2016)
12. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436 (2015)
13. Barker, J.P., Cooke, M.P., Ellis, D.P.: Decoding speech in the presence of other sources. Speech Commun. **45**, 5–25 (2005)
14. Wang, D., Brown, G.J.: Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, Hoboken (2006)
15. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., et al.: CNN architectures for large-scale sound classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135 (2017)
16. Brown, A.L.: Soundscapes and environmental noise management. Noise Control Eng. **58**, 493 (2010)
17. Gelfand, S.A.: Hearing: An Introduction to Psychological and Physiological Acoustics, 6th edn. CRC Press, Boca Rato (2017)
18. Kidd, G.J., Mason, C.R., Richards, V.M., Gallun, F.J., Durlach, N.I.: Informational masking. Audit. Percept. Sound Sources, Springer Handb. Audit. Res. pp. 143–189 (2008)
19. Nilsson, M., Bengtsson, J., Klaeboe, R.: Environmental Methods for Transport Noise Reduction. CRC Press, Boca Raton (2014)
20. Westermann, A., Buchholz, J.M.: The influence of informational masking in reverberant, multi-talker environments. J. Acoust. Soc. **138**, 584–593 (2015)
21. Nilsson, M., et al.: Perceptual effects of noise mitigation. In: Environmental Methods for Transport Noise Reduction, pp. 195–220 (2014)
22. Oldoni, D., et al.: A computational model of auditory attention for use in soundscape research. J. Acoust. Soc. Am. **134**, 852–861 (2013)
23. Kaya, E.M., Elhilali, M.: Modelling auditory attention. Philos. Trans. R. Soc. B Biol., vol. 372 (2017)
24. Kaya, E.M., Elhilali, M.: Investigating bottom-up auditory attention. Front. Hum. Neurosci. **8**, 1–12 (2014)
25. Laffitte, P., Wang, Y., Sodoyer, D., Girin, L.: Assessing the performances of different neural network architectures for the detection of screams and shouts in public transportation. Expert Syst. Appl. **117**, 29–41 (2019)
26. Atrey, P.K., Maddage, N.C., Kankanhalli, M.S.: Sound based event detection for multimedia surveillance. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Vol. 5, (2006)
27. Kong, Q., Xu, Y., Sobieraj, I., Wang, W., Plumbley, M.D.: Sound event detection and time-frequency segmentation from weakly labelled data. IEEE/ACM Trans. Sound Speech Lang. Process. **27**(4), 777–787 (2019)
28. Imoto, K., Ono, N.: Spatial cepstrum as a spatial feature using a distributed microphone array for acoustic scene analysis. IEEE/ACM Trans. Sound Speech Lang. Process. **25**(6), 1335–1343 (2017)

29. Janjua, Z.H., Vecchio, M., Antonini, M., Antonelli, F.: IRESE: an intelligent rare-event detection system using unsupervised learning on the IoT edge. Eng. Appl. Artif. Intell. **84**, 41–50 (2019)
30. Lim, M., et al.: Convolutional neural network based sound event classification. KSII Trans. Internet Inf. Syst. **12**(6) (2018)
31. Kürby, J., Grzeszick, R., Plinge, A., Fink, G.A.: Bag-of-features acoustic event detection for sensor networks. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), pp. 55–59 (2016)
32. Vafeiadis, A., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., Hamzaoui, R.: Sound content analysis for unobtrusive event detection in smart homes. Eng. Appl. Artif. Intell. **89**, 103226 (2020)
33. Kumar, A., Raj, B.: Sound event detection using weakly labeled data. In: Proceedings of the 24th ACM international conference on Multimedia, pp. 1038–1047 (2016)
34. Derakhshan, M., Marvi, H.: Providing an adaptive model with two adjustable parameters for sound event detection and classification in environmental signals. Tabriz J. Electr. Eng. **49**(2), 565–576 (2019)
35. Crockett, B.G., Seefeldt, A.J.: U.S. Patent No. 10,523,169. Washington, DC: U.S. Patent and Trademark Office (2019)
36. Nasiri, A., Cui, Y., Liu, Z., Jin, J., Zhao, Y., Hu, J.: SoundMask: robust sound event detection using mask R-CNN and frame-level classifier. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 485–492. IEEE (2019)
37. Soni, S., Dey, S., Manikandan, M.S.: Automatic sound event recognition schemes for context-aware sound computing devices. In: 2019 Seventh International Conference on Digital Information Processing and Communications (ICDIPC), pp. 23–28. IEEE (2019)
38. Hadi, M., Pakravan, M.R., Razavi, M.M.: An efficient real-time voice activity detection algorithm using teager energy to energy ratio. In: 2019 27th Iranian Conference on Electrical Engineering (ICEE), pp. 1420–1424. IEEE (2019)
39. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 1041–1044 (2014)
40. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. **24**(3), 279–283 (2017)