Learning representations of sound using trainable COPE feature extractors

Strisciuglio, Nicola; Vento, Mario; Petkov, Nicolai

Published in:
Pattern recognition

DOI:
10.1016/j.patcog.2019.03.016

Publication date:
2019

Learning sound representations using trainable COPE feature extractors

Nicola Strisciuglio[a], Mario Vento[b], Nicolai Petkov[a]

[a] *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence - University of Groningen, The Netherlands.*
[b] *Department of Information and Electrical Engineering and Applied Mathematics (DIEM) - University of Salerno, Italy.*

Abstract

Sound analysis research has mainly been focused on speech and music processing. The deployed methodologies are not suitable for analysis of sounds with varying background noise, in many cases with very low signal-to-noise ratio (SNR).

In this paper, we present a method for the detection of patterns of interest in audio signals. We propose novel trainable feature extractors, which we call COPE (Combination of Peaks of Energy). The structure of a COPE feature extractor is determined using a single prototype sound pattern in an automatic configuration process, which is a type of representation learning. We construct a set of COPE feature extractors, configured on a number of training patterns. Then we take their responses to build feature vectors that we use in combination with a classifier to detect and classify patterns of interest in audio signals.

We carried out experiments on four public data sets: MIVIA audio events, MIVIA road events, ESC-10 and TU Dortmund data sets. The results that we achieved (recognition rate equal to 91.71% on the MIVIA audio events, 94% on the MIVIA road events, 81.25% on the ESC-10 and 94.27% on the TU Dortmund) demonstrate the effectiveness of the proposed method and are higher than the ones obtained by other existing approaches. The COPE feature extractors have high robustness to variations of SNR. Real-time performance is

*Corresponding author: n.strisciuglio@rug.nl

achieved even when the value of a large number of features is computed.

Keywords: audio analysis, event detection, peaks of energy, representation learning, trainable feature extractors

---

1. Introduction

Methods and systems for the automatic analysis of people and vehicle behavior, scene understanding, familiar place recognition and human-machine interaction are traditionally based on computer vision techniques. In robotics or public security, for instance, there has been a great effort to equip machines with capabilities for autonomous visual understanding. However, video analysis has some weak points, such as sensitivity to light changes and occlusions, or limitation to the field of view of the camera. Sound is complementary to visual information and can be used to improve the capabilities of machines to deal with the surrounding environment. Furthermore, there are cases in which video analysis cannot be used due to privacy issues (e.g. in public toilets).

In this paper we focus on automatic learning of representations of sounds that are suitable for pattern recognition, in the context of environmental sound analysis for detection and classification of audio events. Recently, the interest in automatic analysis of environmental sounds increased because of various applications in intelligent surveillance and security [1], assistance of eldery people [2], monitoring of smart rooms [3], home and social robotics [4], etc.

A large part of sound analysis research in the past years focused on speech recognition [5], speaker identification [6] and music classification [7]. Features and classifiers for voice analysis are established and widely used in practical systems: spectral or cepstral features in combination with classifiers based on Hidden Markov Models or Gaussian Mixture Models. However, state of the art methods for speech and music analysis do not give good results when applied to environmental sounds, which have highly non-stationary characteristics [8]. Most speech recognition methods assume that speech is based on a phonetic structure, which allows to analyze complex words or phrases by splitting them

in a series of simple phonemes. In the case of environmental sound there is no underlying phoneme-like structure. Moreover, human voice has very specific frequency characteristics that are not present in other kinds of sound. For example, interesting events for surveillance applications, such as gun shots or glass breaking usually have high-frequency components that are not present in speech. For speech recognition and speaker identification the sound source is typically very close to the microphone. It implies that background noise has lower energy than foreground sounds and does not impair considerably the performance of the recognition system. Environmental sound sources can be, instead, at any distance from the microphone. Hence, the background noise can have relatively high energy, so determining very low or even negative signal-to-noise ratio (SNR).

Existing methods for detection of audio events, for which we provide an extensive overview in Section 2, are based on the extraction of hand-crafted features from the audio signal. The features extracted from (a part of) the audio signal are submitted to a classification system. The employed features describe stationary and non-stationary properties of the signals [9]. This approach to pattern recognition requires a feature engineering step that aims at choosing or designing a set of features that describe important characteristics of the sound for the problem at hand. Widely used features are mainly borrowed from the field of speech recognition: responses of log-frequency filters, Mel-frequency cepstral coefficients, wavelet transform coefficients among others. The choice of effective features or combination of them is a critical step to build an effective system and requires considerable domain knowledge.

More recent approaches do not rely on hand-crafted features but rather involve automatic learning of data representations from training samples by using deep learning and convolutional neural networks (CNN) [10]. CNNs were originally proposed for visual data analysis, but have also been successfully applied to speech [11], music processing [12] and sound scene classification [13]. While they achieve very good performance, they require very large amount of labeled training data which is not always available.

In this work, we propose trainable feature extractors for sound analysis which we call COPE (Combination of Peaks of Energy). They are trainable as their structure is not fixed in advance but it is rather learned in an automatic configuration procedure using a single prototype pattern. This automatic configuration of feature extractors is a type of *representation learning*. It allows to automatically construct a suitable data representation to be used together with a classifier and does not require considerable domain expertise. We configure a number of COPE feature extractors on training sounds and use their responses to build a feature vector, which we then employ as input to a classifier. With respect to [14], in which we reported preliminary results obtained using COPE feature extractors on sound events with the same SNR, in this work we provide: *a*) a detailed formulation of the configuration and application steps of COPE features, *b*) a thourough validation of the performance of a classification system based on COPE features when tested with sounds with different values of SNR, *c*) an extension of the MIVIA audio events data set the includes null or negative SNR sound events and *d*) a wide comparison of the proposed method with other existing approaches on four benchmark data sets. Furthermore, we discuss the importance of robustness to variations of the background noise and SNR of the events of interest, for applications of sound event detection in Section 5.4. We provide a detailed analysis of the contribution of the COPE features to the improvement of sound event detection and classification performance with respect to existing approaches.

The design of COPE feature extractors was inspired by certain properties of the inner auditory system, which converts the sound pressure waves that reach our ears into neural stimuli on the auditory nerve. In the Appendix A we provide some details about the biological mechanisms that inspired the design of the COPE feature extractors.

We validate the effectiveness of the proposed COPE feature extractors by carrying out experiments on the following public benchmark data sets: MIVIA audio events [15], MIVIA road events [16], ESC-10 [17], TU-Dortmund [18].

The main contributions of this work are: *a*) novel COPE trainable feature

extractors for representation learning of sounds that are automatically configured on training examples, *b*) a method for audio event detection that uses the proposed features, *c*) the release of an extended version of the MIVIA audio events data set with sounds at null and negative SNR.

The rest of the paper is organized as follows. In Section 2 we review related works, while in Section 3 we present the COPE feature extractors and the architecture of the proposed method. We describe the data sets used for the experiments in Section 4. We report the results that we achieved, a comparison with existing methods and an analysis of the sensitivity of the performance of the proposed method with respect to the parameters of the COPE feature extractors in Section 5. We provide a discussion in Section 6 and, finally, draw conclusions in Section 7.

2. Related works

Representation learning has recently received great attention by researchers in pattern recognition with the aim of constructing reliable features by direct learning from training data. Methods based on deep learning and CNNs were proposed to learn features for several applications: age and gender estimation from facial images [19], action recognition [20], person re-identification [21], hand-written signature verification [22], and also sound analysis [23]. Other approaches for feature learning focused on sparse dictionary learning [24, 25], learning vector quantization [26], and on extensions of the bag of features approach based on neural networks [27] or higher-order pooling [28].

In the context of audio analysis research, it is common to organize existing works on sound event detection according to the feature sets and classification architectures that they employ. Early methods approached the problems of sound event detection and classification by dividing the audio signal into small, partially overlapped frames and computing a feature vector for each frame. The used features ranged from relatively simple (e.g. frame energy, zero-crossing rate, sub-band energy rate) to more complicated ones (e-g. Mel-

frequency Cepstral Coefficients [29], log-frequency filter banks [30], perceptual linear prediction coefficients [31], etc.). The frame-level feature vectors were then used together with a classifier to perform a decision. Gaussian Mixture Model (GMM) based classifiers were largely employed to classify the frames as part of sounds of interest or background [32, 33]. To limit the influence of background sounds on the classification performance, One-Class Support Vector Machines were proposed [34].

Spectro-temporal features based on spectrogram or other time-frequency representations were also developed [35, 36]. Inspired by the way the inner auditory system of humans responds to the frequency of the sounds, an auditory image model (AIM) was proposed [37]. The AIM was used as basis for improved models which are called stabilized auditory images (SAI) [38]. In [39], the event detection was formulated as an object detection problem in a spectrogram-like representation of the sound, and approached by using a cascade of AdaBoost classifiers. The design of hand-crafted features poses some limitations to the construction of systems that are robust to varying conditions of the events of interest and requires considerable domain knowledge.

In order to construct more reliable systems, efforts towards automatic learning of features from training data by means of machine learning techniques were made. Various approaches based on *bag of features* were proposed for sound event representation and classification [40, 41]. A code-book of basic audio features (also called *audio words*) is directly learned from training samples as result of a quantization of the feature space by means of various clustering algorithms (e.g. k-Means or fuzzy k-Means). A comparison of hard and soft quantization of audio words was performed in [15]. Other approaches for the construction of a code-book of basic audio words were also based on non-negative matrix factorization [42] or sparse coding [43]. In the bag of features representation, the information about the temporal arrangement of the audio words is lost. This was taken into account in [44] and [45], where a feature augmentation and a classifier based on Genetic Motif Discovery were proposed, respectively. The sequence of audio words were also employed in [46] and [47]. The temporal information was

described by a pyramidal approach to bag of features in [18, 48]. A method for sound representation learning based on Convolutional Neural Networks (CNN) was proposed in [49]. Learning features from training samples does not require an engineering effort and allows for the adaptation of the recognition systems to various problems. However, the effectiveness and generalization capabilities of learned features depend on the amount of available training data.

Evaluation of algorithms for audio event detection on public benchmark data sets is a valuable tool for objective comparison of performance. The great attention that was dedicated to music and speech analysis determined the publication of several data sets used in scientific challenges for benchmarking of algorithms. The MIREX challenge series evaluated systems for music information retrieval (MIR) [50]. The CHiME challenge focused on speech analysis in noisy environments [51]. The "Acoustic event detection and classification" task of the CLEAR challenges (2006 and 2007) focused on the detection of sound events related to seminars, such as speech, chair moving, door opening and applause [52]. Recently, the DCASE challenge [53] stimulated the interest of researchers on audio processing for the analysis of environmental sounds. The attention was driven towards audio event detection and classification and scene classification.

3. Method

In Figure 1, we show an overview of the architecture of the proposed method. The algorithm is divided in two phases: configuration and application.

In the configuration phase (dashed line), the Gammatonegrams (see details in Section 3.1) of prototype training sounds are used to configure a set of COPE feature extractors (see Section 3.2.2). Successively, the response of the set of COPE feature extractors, computed on the sounds in the training set, are employed to construct COPE feature vectors (Figure 1b-d). A multi-class SVM classifier is finally trained using the COPE feature vectors (Figure 1e) to distinguish between the classes of interest for the application at end.

In the application phase, the previously configured set of COPE feature extractors is applied to extract feature vectors from input unknown sounds and the multi-class SVM classifier is used to detect and classify sound events of interest. The implementation of the COPE feature extractors and the proposed classification architecture is publicly available[1].

3.1. Gammatonegram

The traditional and most used time-frequency representation of sounds is the spectrogram, in which the energy distribution over frequencies is computed by dividing the frequency axis into sub-bands with equal bandwidth. In the human auditory system, the resolution in the perception of differences in frequency changes according to the base frequency of the sound. At low frequency the band-pass filters have a narrower bandwidth than the ones at high frequency. This implies higher time resolution of filters at high frequency that are able to better catch high variations of the signal. In this work we employ a bank of Gammatone band-pass filters, whose bandwidth increases with increasing central frequency. The functional form of Gammatone is biologically-inspired and models the response of the cochlea membrane in the inner ear of the human auditory system [54].

The impulse response of a Gammatone filter is the product of a statistical distribution called *Gamma* and a sinusoidal carrier tone. It is formally defined as:

$$h_i(t) = \begin{cases} at^{n-1}e^{-2\pi B_i t}cos(2\pi\omega_i t + \phi), & t \geq 0 \\ 0, & else \end{cases}, \tag{1}$$

where ω_i is the central frequency of the filter, and ϕ is its phase. The constant a controls the gain and n is the order of the filter. The parameter B_i is a decay factor and determines the bandwidth of the band-pass filter. The relation between the central frequency of a Gammatone filter and its bandwidth is given

[1]The code is available at http://gitlab.com/nicstrisc/COPE

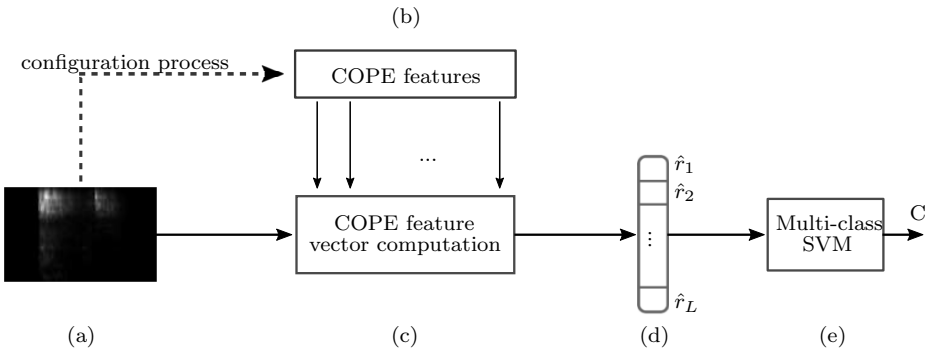


Figure 1: Architecture of the proposed method. The (a) Gammatonegram of the training audio samples is computed in the training phase (dashed arrow), and used to configure a (b) set of COPE feature extractors. The learned features are used in the application phase to (c) process the input sound and (d) construct feature vectors with their responses. A (e) multi-class SVM classifier is, finally, employed to detect events of interest.

by the Equivalent Rectangular Bandwidth (ERB):

$$B_i = \left[\left(\frac{\omega_i}{Q_{ear}} \right)^p + (B_{min})^p \right]^{1/p} \tag{2}$$

where Q_{ear} is the asymptotic filter quality at high frequencies and B_{min} is the minimum bandwidth at low frequencies, while p is usually equal to 1 or 2. In [55], the parameters $Q_{ear} = 9.26779$, $B_{min} = 24.7$ and $p = 1$ where determined by measurements from notched-noise data. In Figure 2a, we show the impulse response of two Gammatone filters with low ($\omega_1 = 115.1$ Hz) and higher ($\omega_2 = 1.96$ KHz) central frequencies. The filter with higher central frequency has larger bandwidth, as it can be seen from their frequency response in Figure 2b.

We filter the input signal $x(t)$ with a bank of Γ Gammatone filters $\underline{\mathbf{h}}(t) = [h_1(t), h_2(t), \ldots, h_\Gamma(t)]^T$. The response of the i-th filter to an input signal $x(t)$ is the convolution of the input signal with the impulse response $h_i(t)$:

$$\tilde{x}_i(t) = x(t) * h_i(t). \tag{3}$$

We divide the input audio signal in frames of F samples and process every frame by a bank of Gammatone filters in order to capture the short-time

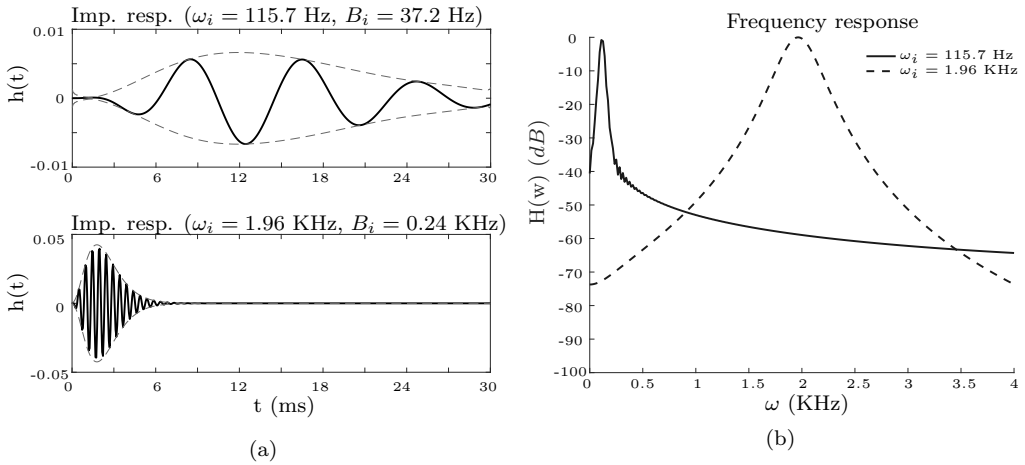


Figure 2: (a) Impulse responses of two Gammatone filters (with central frequencies $\omega_1 = 115.1$ Hz and $\omega_2 = 1.96$ KHz). The dashed lines represent the envelope of the sinusoidal tone. The (b) frequency responses of the filters in (a): the filter with higher central frequency (dashed line) has larger bandwidth ($B_2 = 240$ Hz while $B_1 = 37.2$ Hz).

properties of the energy distribution of the sound. Two consecutive frames have $F/2$ samples in common, which means that they overlap for 50% of their length. This ensures continuity of analysis and that border effects are avoided. Given an input signal with N samples, the number of concerned frames is $\Theta = \lfloor 2(N - F)/F \rfloor + 1$. We finally construct the Gammatonegram of a sound as a matrix $\mathcal{X} \in \mathcal{R}^{\Gamma \times \Theta}$, whose j-th column corresponds to $[\tilde{x}_i(jF/2), \tilde{x}_i(jF/2 + 1), \ldots, \tilde{x}_i(jF/2 + F - 1)]^T$ with $j = 0, 1, \ldots, \Theta - 1$. The energy value of the i-th frequency channel in the Gammatonegram at the j-th time instant is:

$$\mathcal{X}_{i,j} = \sqrt{\frac{1}{F} \sum_{k=0}^{F-1} \left[\tilde{x}_i \left(j\frac{F}{2} + k \right) \right]^2}. \quad (4)$$

In Figure 3a, we show the Gammatonegram representation of a sample scream sound. It is similar to the spectrogram, with the substantial difference that the frequency axis has a logarithmic scale and the bandwidth of the band-pass filters increases linearly with the value of the central frequency.

3.2. COPE features

The configuration and application of COPE feature extractors involve a number of steps that we explain in the following of this section. In the application phase, given the Gammatonegram representation of a sound, a COPE feature extractor responds strongly to patterns similar to the one used in the configuration step. It also accounts for some tolerance in the detection of the pattern of interest, so being robust to distortions due to noise or to varying SNR.

3.2.1. Local energy peaks

The energy peaks (local maxima) in a Gammatonegram $\mathcal{X}$ are highly robust to additive noise [56]. This property provides underlying robustness of the designed COPE features to variation of the SNR of the sounds of interest. We consider that a point is a peak if it has higher energy than the points in its neighborhood. We suppress non-maxima points in the Gammatonegram and obtain an energy peak response map, as follows:

$$\mathcal{P}_{\mathcal{X}}(t,f) = \max_{\substack{t-\Delta t \leq t' \leq t+\Delta t \\ f-\Delta f \leq f' \leq f+\Delta f}} \mathcal{X}(t',f') \tag{5}$$

where $t = 0, \ldots, \Theta - 1$ and $f = 0, \ldots, \Gamma - 1$. The values Δt and Δf determine the size, in terms of time and frequency, of the neighborhood around a time-frequency point in which the local energy is evaluated (in this work we consider 8-connected pixels). We consider the arrangement (hereinafter constellation) of a set of such time-frequency points as a description of the distribution of the energy of a particular sound.

3.2.2. Configuration of a COPE feature extractor

Given the constellation of energy peaks of a sound and a reference point (in our case the point that correspond to the highest peak of energy), we determine the structure of a COPE feature extractor in an automatic configuration process. For the configuration one has to set the support size of the COPE feature extractor, i.e. the size of the time interval around the reference point in which to consider energy peaks.

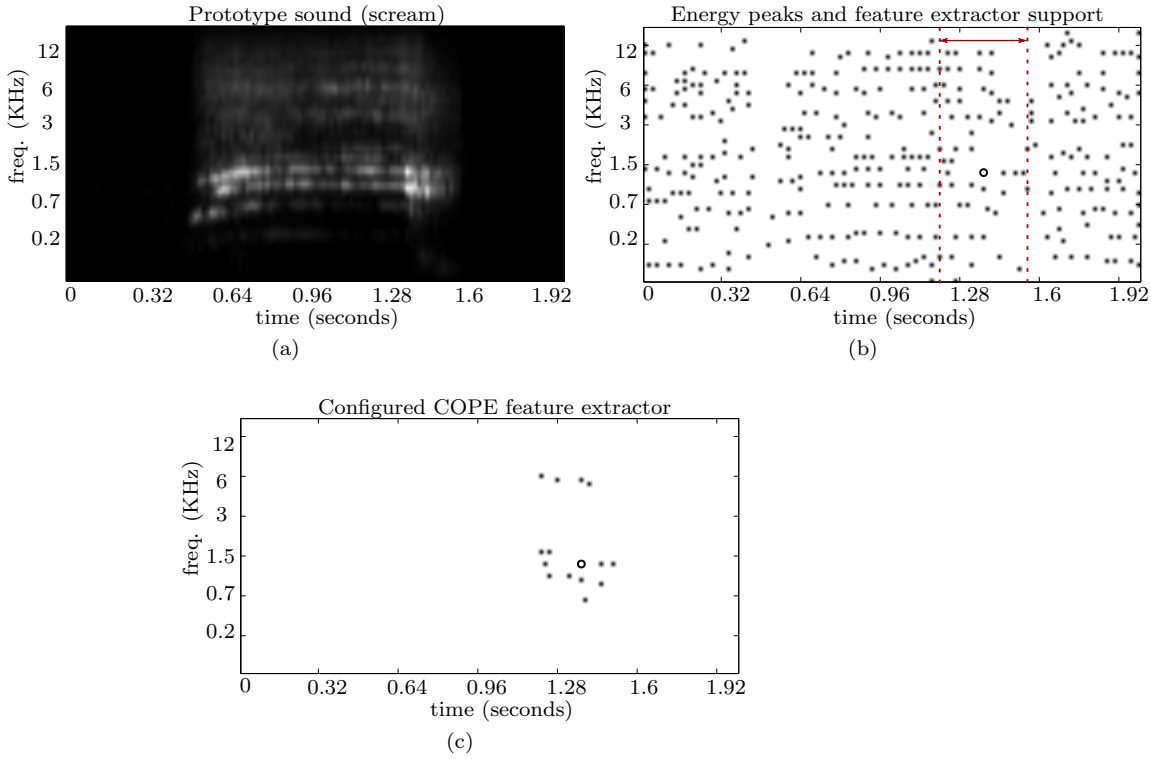


Figure 3: Example of configuration of a COPE feature extractor performed on the (a) Gammatonegram representation of a scream. The (b) energy peaks are extracted and a support (dashed lines) is chosen around a reference point (small circle). The (c) configured feature extractor is composed of only those points whose energy is higher than a fraction t_1 of the energy of the reference point.

In Figure 3, we show an example of the configuration process on the scream sound in Figure 3a. First, we find the position of the local energy peaks and select a reference point (small circle in Figure 3b) around which we define the support size of the feature extractor. The support is contained between the two dashed (red) lines in Figure 3b. We consider the positions of only those peaks that fall within the support of the feature extractor and whose energy is higher than a fraction t_1 of the highest peak of energy (Figure 3c). Every peak point p_i is represented by a 3-tuple $(\Delta t_i, f_i, e_i)$: Δt_i is the temporal offset with respect to the reference point, f_i is its corresponding frequency channel in the

Gammatone filterbank and e_i is the value of the energy contained in it.

The configuration process results in a set of tuples that describe the constellation of energy peaks in the Gammatonegram image of a sound. We denote by $S = \{(\Delta t_i, f_i, e_i) \mid i = 1, \ldots, L\}$ the set of 3-tuples, where L is the number of considered peaks within the support of the filter.

3.2.3. Feature computation

Given a Gammatonegram, we compute the response of a COPE feature extractor as a combination of its weighted and shifted energy peaks. We define the weighting and shifting of the i-th energy peak as :

$$s_i(t) = \max_{t',f'} \{\psi(t - \Delta t_i - t', f_i - \Delta f_i - f') G_{\sigma'}(t', f')\} \tag{6}$$

where $-3\sigma' \leq t', f' \leq 3\sigma'$.

The function $\psi(t, f)$ can be seen as a response map of the similarity between the detected energy peak in the input Gammatonegram and the corresponding one in the model. In this work we consider $\psi(t, f) = \mathcal{P}_{\mathcal{X}}(t, f)$, so as to account only for the position and energy content of the peak points in the constellation. We weigth the response $\psi(t, f)$ with a Gaussian weighting function $G_{\sigma'}(\cdot, \cdot)$ that allows for some tolerance in the expected position of the peak points. This choice is supported by evidence in the auditory system that vibrations of the cochlea membrane due to a sound wave of a certain frequency excite neurons specifically tuned for that frequency and also neighbor neurons [57]. The size of the tolerance region is determined by the standard deviation σ' of the function $G_{\sigma'}$, which is a parameter and we set as $\sigma' = \sigma_0/2$.

The value of a COPE feature is computed with a sliding window that shifts on the Gammatonegram of the input sound. Formally, we define it as the geometric mean of the weighted and shifted energy peak responses in Eq. 6:

$$r(t) = \left| \left(\prod_{i=1}^{|S|} s_i(t) \right)^{1/|S|} \right|_{t_2}, \tag{7}$$

where t_2 is a threshold value. Here, we set $t_2 = 0$, so to not suppress any response. The value of a COPE feature for a sound in an interval delimited by

$[T_1, T_2]$ is given by max-pooling of the response $r(t)$ with $T_1 \leq t \leq T_2$:

$$\hat{r} = \max_{t \in [T_1, T_2]} r_i(t) \tag{8}$$

3.3. COPE feature vector

We configure a set of COPE feature extractors on K training audio samples from different classes. For a given interval of sounds $[T_1, T_2]$, we then construct a feature vector as follows:

$$\mathbf{v}_{[T_1,T_2]} = [\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_K]\,. \tag{9}$$

3.4. Classifier

We use the COPE feature vectors to train a classifier, which is able to assign the input sound to one of the M classes of interest. The COPE feature vectors are not dependent on a specific classier and thus one can employ them together with any multi-class classifier.

In this work, we employ a multi-class SVM classifier, designed according to a *one-vs-all* scheme, in which M binary SVM classifiers (where M corresponds to the number of classes) are trained to recognize samples from the classes of interest. We use linear SVMs with soft-margin as they provide already satisfactory results and are easy to train. We set the hyperparameter $c = 1$ for the training of each SVM, which indicates the trade-off between training error and size of the classification margin while training the SVM classifier (see [58] for reference). We train the i-th SVM ($i = 0, \ldots, M-1$) by using as positive examples those of the class C_i and as negative samples those of all the remaining classes. In this scheme, the training of each SVM classifier is an unbalanced problem, as the cardinality of the samples from the negative class $|N|$ outnumbers that of the samples from the positive class $|P|$. We thus employ an implementation of the SVM algorithm that includes a cost-factor $J = |N|/|P|$ by which training errors on positive examples outweight errors on negative examples[2] [59]. In this

[2] Available in the SVMlight library - http://svmlight.joachims.org/

way, the training errors for the positive and negative examples have the same influence in the overall optimization.

During the test phase, each SVM classifier assigns a score m_i to the given sample under test (i.e. a COPE feature vector that represents the sound to classify). We analyze the SVM scores m_i and assign to the test vector the class that corresponds to the SVM that gives the highest classification score. We assign the sample under test to the reject class C_0 (background sound) in case all the scores are negative. We formally define the classification rule as:

$$C = \begin{cases} C_0, & if\ m_i < 0 \qquad \forall i = 0, \ldots, M-1 \\ \arg\max_i m_i, & else. \end{cases} \tag{10}$$

4. Data sets

We carried out experiments on four public data sets, namely the MIVIA audio events [15], MIVIA road events [16], ESC-10 [17] and TU-Dortmund [18] data sets.

4.1. MIVIA audio events

Typical sounds of interest for intelligent surveillance applications are glass breakings, gun shots and screams. In the MIVIA audio events data set, such sounds are superimposed to various background sounds and have different SNRs ($\{5, 10, \ldots, 30\}dB$). This simulates the occurrence of sounds in different environments and at various distances from the microphone. We extended the data set by including cases in which the energy of the sounds of interest is equal or lower than the one of the background sound, so having null or negative SNR. Thus, adopting the same procedure described in [15], we created two versions of the audio events at $0dB$ and $-5dB$ SNR. The final data set[3] contains a total of 8000 events for each class, divided into 5600 events for training and 2400 events for testing equally distributed over the considered values of SNR. The

[3]The data set is publicly available at the url *http://www.gitlab.com/nicstrisc/COPE*

Table 1: Details of the composition of the MIVIA audio events data set. The total duration of the sounds is expressed in seconds.

MIVIA audio events data set

	Training set		**Test set**	
	#Events	**Duration (s)**	**#Events**	**Duration (s)**
BN	-	77828.8	-	33382.4
GB	5600	8033.1	2400	3415.6
GS	5600	2511.5	2400	991.3
S	5600	7318.4	2400	3260.5

audio clips are PCM sampled at 32KHz with a resolution of 16 bits per sample. Hereinafter we refer at glass breaking with *GB*, at gun shots with *GS* and at screams with *S*. We indicate the background sound with *BN*. In Table 1, we report the details of the composition of the extended data set.

4.2. MIVIA road events

The MIVIA road events data set contains car crash and tire skidding events mixed with typical road background sounds such as traffic jam, passing vehicles, crowds, etc. A total of 400 sound events (200 car crashes and 200 tire skiddings) are superimposed to various road sounds ranging from very quiet (e.g. in country roads) to highly noisy traffic jams (e.g. in the center of a big city) and highways. The sounds of interest are distributed over 57 audio clips of about one minute each, which are organized into four independent folds (in each fold 50 events per class are present) for cross-validation experiments. The audio signals are sampled at 32KHz with a resolution of 16 bits per PCM sample. In the rest of the paper, we refer at car crash with *CC* and at tire skidding with *TS*.

4.3. ESC-10

The ESC-10 data set is composed of 400 sounds divided in ten classes (*dog bark, rain, sea waves, baby cry, clock tick, sneeze, helicopter, chainsaw, rooster, fire crackling*), each of them containing 40 samples. The sounds are sampled at 44.1 KHz with a bit rate of 192 kbit/s and their total duration is about

33 minutes. The data set is organized in five independent folds. The average classification accuracy achieved by human listeners is 95.7%.

4.4. TU Dortmund

The TU Dortmund data set was recorded in a smart room with a microphone embedded on a table. The data set is composed of sounds from eleven classes (*chair, cup, door, keyboard, laptop keys, paper sheets, pouring, rolling, silence, speech, steps*), divided in a training and a test sets. The sounds of interest are sampled at 48 Khz and are mixed with the background sound of the smart room. A ground truth with the start and end points of the sounds is provided. We constructed a second observer ground truth, which contains a finer grain manual segmentation of the events.

5. Experiments

5.1. Performance evaluation

For the MIVIA audio events and the MIVIA road events data sets we adopted the experimental protocol defined in [15]. The performance evaluation is based on the use of a time window of T_w seconds that forward shifts on the audio signal by ΔT_w seconds. An event is considered correctly detected if it is detected in at least one of the time windows that overlap with it. Besides the recognition rate and confusion matrix, we consider two types of error that are important for performance evaluation: the detection of events of interest when only background sound is present (false positive) and the case when an event of interest occurs but it is not detected (missed detection). In case a false positive is detected in two consecutive time windows, only one error is counted. We measured the performance of the proposed method by computing the recognition rate (RR), false positive rate (FPR), error rate (ER) and miss detection rate (MDR). Moreover, in addition to the receiver operating characteristic (ROC) curve and in order to assess the overall performance of the proposed method we compute the Detection Error Trade-off (DET) curve. It is a plot of the trade-off

Table 2: Classification matrix obtained by the proposed method on the extended MIVIA audio events data set. GB, GS and S indicate the classes in the data set (see Section 4.1), while MDR is the miss detection rate.

Results - MIVIA audio events data set

		Detected class			
		GB	**GS**	**S**	**MDR**
True class	**GB**	95.33%	2.13%	1.25%	1.29%
	GS	4.33%	89.25%	2.58%	3.83%
	S	1.5%	4.92%	87.79%	5.79%

between the false positive rate and the miss detection rate and gives an insight of the performance of a classifier in terms of its errors. In contrast to the ROC curve, in the DET curve the axis are logarithmic in order to highlight differences between classifiers in the critical operating region. The closer the curve to the point $(0, 0)$, the better the performance of the system.

For the ESC-10 and TU Dortmund data sets, we evaluate the performance for the classification of isolated audio events. This type of evaluation is done according to the structure of these data sets and to make possible a comparison with the results achieved by other approaches. We compute the average recognition rate (RR) and the F-Measure $F = 2RePr/(Re + Pr)$, where $Pr = TP/(TP + FP)$ is the precision and $Re = TP/(TP + FN)$ is the recall. TP, FP and FN are the number of true positive, false positive and false negative classifications, respectively. In the case of the MIVIA road events and the ESC-10 data sets, we perform cross-validation experiments.

5.2. Results

In Table 2, we report the classification matrix that we obtained on the extended version of the MIVIA audio events data set. The average recognition rate for the three classes is 90.7%, while the miss detection rate and the error rate are 3.7% and 5.6%, respectively. We obtained an FPR equal to 7.1%, of which 1.25% are glass breakings, 2.74% are gun shots and 3.11 are screams.

In Table 3 we report the classification matrix achieved by the proposed

Table 3: Average results obtained by the proposed method on the MIVIA road events data set. CC and TS are acronyms for the classes in the data set (see Section 4.2).

Results - MIVIA road events

		Guessed class		
		CC	**TS**	**MDR**
True class	**CC**	92%	2%	6%
	TS	0.5%	96%	3.5%

approach on the MIVIA road events data set. The average RR is 94% with a standard deviation of 4.32%, while the average FPR is 3.94% with a standard deviation of 1.82%. The results are in line with the ones achieved on the MIVIA audio events data set. The low standard deviation of the recognition rate is indicative of good generalization capabilities.

The proposed method shows high performance on the ESC-10 and TU Dortmund data sets, which both contain a larger number of classes than in the MIVIA data sets, but with a lower number of samples per class. We achieved $RR = 81.25\% \quad (5.38\%), Pr = 0.8263 \quad (0.053), Re = 0.8125 \quad (0.054), F = 0.8048 \quad (0.059)$ on the ESC-10 data set (the standard deviation of each measure is in brackets). On the TU Dortmund data set we achieved $RR = 94.27\%, Pr = 0.9479, Re = 0.9519$ and $F = 0.9469$. In the following, we compare the achieved results with the ones reported in other works.

5.3. Results comparison

In Table 4, we report the results that we achieved on the MIVIA audio event data set, compared with the ones of existing methods. In the upper part of the table we compare the results achieved by considering the classification of sound events with positive SNR only. In the lower part of the Table, we report the results achieved by including also sound events with negative and null SNR in the evaluation.

It is important to clarify that the methods described in [15, 60] employ the same multi-class *one-vs-all* linear SVM classifiers of this work. The results that we report using SoundNet features [61] were obtained by using the features com-

Table 4: Comparison of the results with the ones of existing approaches on the MIVIA audio events data set. RR, MDR, ER and FPR refer to the metrics described in Section 5.1.

Result comparison - MIVIA audio events data set

Method	**RR**	**MDR**	**ER**	**FPR**
Test with $SNR > 0$				
COPE	96%	**3.1**%	**0.9**%	4.3%
bof_h [15]	84.8%	12.5%	2.7%	2.1%
bof_s [15]	86.7%	10.7%	2.6%	3.1%
Gammatone [60]	88.6%	9.65%	1.4%	**1.4**%
UDWT [60]	77.81%	10.65%	11.54%	6.6%
SoundNet [61]	93.33%	0.67%	6%	22.34%
HRNN [62]	**96.55**%	–	–	–
Test with $SNR > 0$ and $SNR \leq 0$				
COPE	**91.7**%	**2.61**%	**5.68**%	9.2%
bof_h [15]	56.07%	36.43%	7.5%	**5.3**%
bof_s [15]	59.11%	32.97%	7.92%	**5.3**%
SoundNet [61]	84.13%	4%	11.88%	25.9%

puted at the last convolutional layer of the SoundNet network in combination with the same classifier of this work.

SoundNet features obtained comparable recognition rate to the one achieved by the proposed approach, but a considerably higher FPR. The recognition rate achieved by the Hierarchical Recurrent Neural Network classifier (HRNN) proposed in [62] is slightly higher than the ones we obtained, though the HRNN-based approach has more complex design and training procedure, and a different classifier than SVM. The values of MDR, ER and FPR are not reported in [62].

The performance of the proposed method demonstrated high robustness of the COPE feature extractors w.r.t. variations of the SNR. Conversely, for the methods proposed in [15], the performance of the classification systems strongly depend on the SNR of the training sound events. When sounds with only positive SNR are used for training, the recognition rate achieved by the proposed method is almost 10% higher than the one obtained by the approaches proposed in [15, 60]. The performance results of the latter methods decrease strongly

Table 5: Comparison of the results achieved on the MIVIA roads events data set with respect to the methods proposed in [16, 63]. RR, MDR, ER and FPR refer to the evaluation metrics described in Section 5.1.

Comparison of results on MIVIA road events data set

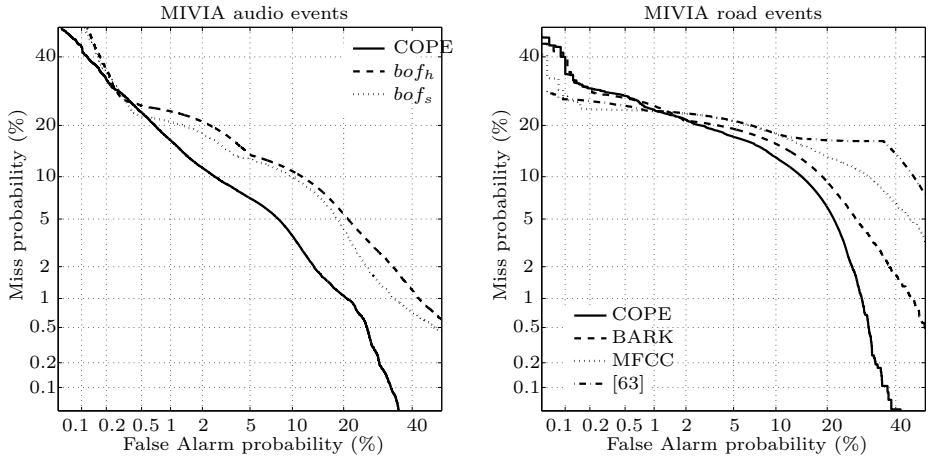
	RR	**MR**	**ER**	**FPR**
COPE	94%	4.75%	1.25%	3.95%
σ	4.32	4.92	1.26	1.82
bof_{BARK} [16]	80.25%	21.75%	3.25%	10.96%
σ	7.75	8.96	2.5	8.43
bof_{MFCC} [16]	80.25%	19%	0.75%	7.69%
σ	11.64	11.63	0.96	5.92
bof [63, 16]	82%	17.75%	0.25%	2.85%
σ	7.79	8.06	1	2.52

(recognition rate more than 30% lower than the one of the proposed method) when sounds with negative SNR are included in the model. We provide an extensive analysis of robustness to variations of SNR in Section 5.4.

In Table 5, we compare the results we obtained on the MIVIA road events data set with the ones reported in [16], where different sets of audio features (BARK, MFCC and a combination of temporal and spectral features) have been employed as short-time descriptors of sounds. We obtained an average recognition rate (94%) that is more than 10% higher than the ones achieved by existing methods, with a lower standard deviation.

In Figure 4a and 4b, we plot the DET curves obtained by our method (solid lines) on the MIVIA audio events and MIVIA road events data sets, respectively, and those of the methods proposed in [15] and [16] (dashed lines). The curve of our method is closer to the point $(0, 0)$ than the ones of other approaches, so confirming higher performance with respect to existing methods on the concerned data sets.

We compare the results that we achieved on the ESC-10 data set with the ones reported by existing approaches in Table 6. The sign '$-$' indicates that the concerned value is not reported in published papers. The highest recognition rate is achieved by SoundNet [61], which is a deep neural network trained on a



(a) (b)

Figure 4: Detection Error Trade-off curves achieved by the proposed method (solid line) compared to the curves achieved by existing methods (dashed lines) on the (a) MIVIA audio events and (b) MIVIA road events data sets. (Notice the logarithmic scales.)

Table 6: Comparison of the results on the ESC-10 data set. In brackets we report the standard deviation of the average performance metrics. RR, MDR, ER and FPR refer to the evaluation metrics described in Section 5.1.

Result comparison on ESC-10 data set

Method	**RR**	**F**
COPE	81.25% (5.38)	0.81 (0.06)
Baseline [17]	66.74% (6.11)	–
Random Forest [17]	72.75% (8.68)	0.72 (0.09)
Piczak CNN [49]	80.25% (5.48)	0.80 (0.06)
Conv. Autoenc. [61]	74.3%	–
Hertel CNN [64]	89.9%	–
SoundNet [61]	92.2%	–
MCLNN. [65]	85.5%	–

very large data set of audio-visual correspondences. Approaches based on CNNs ([49, 61, 64, 65]) are trained with data augmentation techniques and generally perform better than the proposed approach on the ESC-10 data set, which is instead trained only on the original sounds in the ESC-10 data set.

Table 7: Comparison of the results on the TU Dortmund data set. The results were computed with respect to the second observer ground truth that we constructed. RR, MDR, ER and FPR refer to the evaluation metrics described in Section 5.1.

Result comparison on TU Dortmund data set

	RR	**Pr**	**Re**	**F**
COPE	**94.27**%	**94.79**%	**95.19**%	**94.69**%
BoF [18]	90.05%	92.39%	88.82%	90.57%
P-BoF [18]	89.94%	92.24%	88.67%	90.42%
BoSF [18]	90.31%	92.73%	88.13%	90%

In Table 7, we report the results that we achieved on the TU Dortmund data set together with those reported in [18], where a classifier based on bag of features was proposed. Besides the traditional bag of features (BoF) scheme, the authors proposed a pyramidal approach (P-BoF) and the use of super-frames (BoSF) for embedding temporal information about the sequence of features. The results in Table 7 are computed according to the ground truth that we constructed based on a fine segmentation of sounds of interest and that we made publicly available. It is worth noting that the performance results of our method refer to the classification of sound events. For the methods proposed in [18] the evaluation is performed by considering the classification of sound frames.

5.4. Robustness to background noise and SNR variations

We carried out a detailed analysis of the performance of the COPE feature extractors on sounds with different levels of SNR. We trained the proposed classifier following two different schemes. For the first training scheme (that we refer at as T1) we included in the training process only sound events with positive SNR. For the second training scheme (that we refer at as T2) we trained the classifier with all the sounds in the data set, including those with null and negative SNR. In Table 8, we report the results that we achieved for the classification of sounds in the MIVIA audio event data set by training the system according to T1 and T2. We tested both trained models on the whole test set of the

MIVIA audio event data set (including negative and null SNRs). The proposed method showed stronger robustness to changing SNR w.r.t. previously published approaches in [15], especially when samples with null and negative SNR are not included in the training process. This demonstrate high generalization capabilities of the proposed COPE feature extractors to sound events corrupted by high-energy noise.

In Figure 5, we plot the ROC curves relative to the performance achieved at the different levels of SNR of the sounds of interest contained in the MIVIA audio event data set. We observed substantial stability of performance when the sounds of interest have positive (also very low) SNR. The high robustness of the COPE feature extractors with respect to variations of the SNR is attributable to the use of the local energy peaks extracted from the Gammatonegram, which are robust to additive noise. The slightly lower results at negative SNR are mainly due to the changes of the expected energy peak locations caused by high energy of the background sounds. In such cases, most of the wrong classifications are due to errors rather than to miss detection of sounds of interest.

5.5. Sensitivity analysis

We analyzed the sensitivity of the COPE feature extractors with respect to the parameter σ_0 which regulates the degree of tolerance to changes of the

Table 8: Analysis and comparison of stability of results w.r.t. varying value of SNR of the events of interest. Details on the training schemes T1 and T2 are provided in Section 5.4. RR, MDR, ER and FPR refer to the evaluation metrics described in Section 5.1.

Comparison of results on the MIVIA audio events data set

		Training T1				**Training T2**			
Test	**Method**	**RR**	**MDR**	**ER**	**FPR**	**RR**	**MDR**	**ER**	**FPR**
all SNR	COPE	**91.7**%	**2.61**%	**5.68**%	9.2%	**90.7**%	**3.7**%	**5.6**%	7.2%
	bof_h [15]	76.4%	11.64%	11.96%	**5.9**%	56.07%	36.43%	7.5%	**5.3**%
	bof_s [15]	77.81%	10.65%	11.54%	6.6%	59.11%	32.97%	7.92%	**5.3**%
SNR>0	COPE	96%	3.1%	0.9%	4.3%	95.2%	4%	0.8%	2.2%
	bof_h [15]	84.8%	12.5%	2.7%	2.1%	64.63%	31%	4.4%	4.2%
	bof_s [15]	86.7%	10.7%	2.6%	3.1%	68.74%	26.4%	4.9%	4.5%

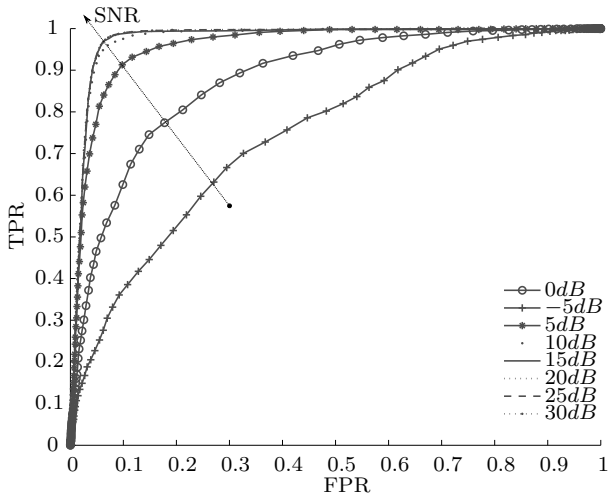


Figure 5: ROC curves obtained by the proposed method on the MIVIA audio events data set at different SNR values ($\{-5, 0, 5, \ldots, 30\}dB$). The arrow indicates increasing values of SNR.

sounds of interest due to background noise or distortion. We used a version of the MIVIA audio events data set specifically built for cross-validation experiments. The data set was released in [15], employing the same procedure used for the MIVIA audio events data set, ensuring statistical independence and high variability among folds. The sound events were divided into $k = 5$ independent folds, each of them containing 200 events of interest per class (times 8 versions of the SNR, as in the original MIVIA events data set). In our analysis, we estimated the variance of the generalization error using the Nadeau-Bengio variance estimator [66], which takes into account the variability of the training and test sets used in cross-validation experiments.

For the configuration of a COPE feature extractor, the user has to choose the size of its support, i.e. the length of the time interval around the reference point in which energy peaks are considered for the configuration (see Section 3.2.2). We experimentally observed that different sizes of the support, namely $s_t = \{200, 300, 400\}$ ms, do not significantly influence the performance of the proposed system on the MIVIA data sets. We report results achieved with a support of 200 ms, which involves a limited number of energy peaks in

Table 9: Sensitivity of the COPE feature extractors to various values the parameter σ_0. Higher the value of σ_0, larger the tolerance of the feature extractor to variations of the pattern of interest. For the generalization error (ER) of cross-validation experiments, we report the value of the Nadeau-Bengio estimator of variance that takes into account the variability in the training and test sets [66]

Sensitivity of COPE feature extractors

	MIVIA audio events				**MIVIA road events**			
$\boldsymbol{\sigma_0}$	$\boldsymbol{ER}$	$\boldsymbol{\hat{\sigma}_{ER}}$	$\boldsymbol{FPR}$	$\boldsymbol{\sigma_{FPR}}$	$\boldsymbol{ER}$	$\sigma_{\hat{ER}}$	$\boldsymbol{FPR}$	$\boldsymbol{\sigma_{FPR}}$
1	35.98%	8.3	19.21%	7.29	28.5%	6.47	17.09%	7.66
2	26.6%	3.92	18.75%	3.16	19%	3.87	21.46%	11.14
3	13.84%	1.41	13.97%	2.85	4.75%	4.41	18.78%	12.5
4	13.83%	2.1	11.53%	2.4	4.75%	3.08	7.44%	4.39
5	14.70%	2.22	9.91%	2.2	6%	3.05	3.94%	1.82
6	14.37%	3.04	10.23%	2.1	6.25%	3.39	3.94%	1.13

the configuration of the feature extractors. One could however choose $s_t = 400$ ms, achieving similar performance to the case in which $s_t = 200$ ms. The drawback is the need of computing and combining the responses of a higher number of energy peaks, which increase the processing time of each feature extractor.

In Table 9, we report the generalization error (ER) and the false positive rate (FPR) as the parameter σ_0 varies. The performance of the proposed system is slightly sensitive to varying values of the parameter σ_0, mostly when they are kept very low. For higher values ($\sigma_0 = 3, 4, 5, 6$), the performance shows more stability. Higher tolerance for the detection of the energy peak positions determines stronger robustness to background noise. It is worth pointing out that too large values of tolerance might cause a loss in the selectivity and descriptive power of the COPE feature extractors and consequently a decrease of the classification performance.

6. Discussion

The high recognition capabilities of the proposed method are attributable to the trainable character and the versatility of the COPE feature extractors. The concept of trainable filters has been previously introduced for visual pattern

recognition. COSFIRE filters were proposed for contour detection [67], key-
point and object detection [68], retinal vessel segmentation [69, 70], curvilinear structure delineation [71, 72, 73], and action recognition [74]. In this work, we extended the concept of trainable feature extractors to sound recognition. It is noteworthy that the proposed COPE feature extractors do not relate with template matching techniques, which are sensitive to variations with respect to the reference pattern. The tolerance introduced in the application phase allows also for the detection of modified versions of the prototype pattern, mainly due to noise or distortion.

An important advantage of using COPE feature extractors is the possibility of avoiding the process of feature engineering, which is a time-consuming task and requires substantial domain knowledge. In traditional sound recognition approaches, hand-crafted features (e.g. MFCC, spectral and temporal features, Wavelets and so on) are usually chosen and combined together to form a feature vector that describes particular characteristics of the audio signals. On the contrary, the automatic configuration of COPE feature extractors consists in learning data representations directly from the sounds of interest. Manual engineering of features is indeed not required.

Representation learning is typical of recent machine learning methods based on deep and convolutional neural networks, which require large amount of training data. When large data sets are not available, new synthetic data is generated by transformations of the original training data. To this concern, the COPE algorithm differs from deep and convolutional neural networks approaches, as it requires only one prototype pattern to configure a new feature. Moreover, the tolerance introduced in the application phase guarantees, to a certain extent, good generalization properties. Because of their flexibility, COPE feature extractors can be thus employed in various sound processing applications such as music analysis [75, 76] or audio fingerprinting [77], among others.

The COPE feature extractors are robust to variations of the background noise and of the SNR of the sounds of interest. In Figure 6a we show the Gammatonegram of a glass breaking sound with SNR equal to $30 dB$. As an

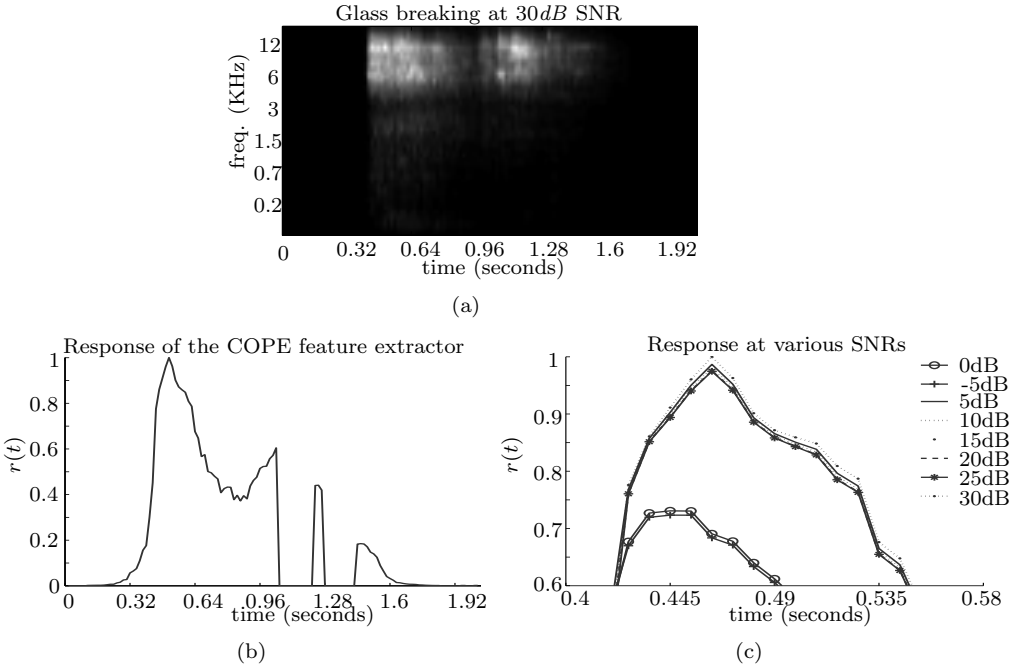


Figure 6: The (a) Gammatonegram of a prototype glass breaking sound used for the configuration of a COPE feature extractor. The (b) response $r(t)$ of the feature extractor computed on the sound in (a). The (c) time-zoomed (between 0.4s and 0.58s) response $r(t)$ at different SNRs ($\{-5, 0, \ldots, 30\} dB$). The response is stable for positive values of SNR and decreases for null or negative SNR values.

example, we configure a COPE feature extractor on this sound and compute its response $r(t)$ on the sound of Figure 6a, which we show in Figure 6b. One can observe that the response is maximum in the same point used as reference point in the configuration phase. The response is null when at least one of the expected energy peaks is not present. In Figure 6c, we show a time-zoomed detail of the response of the feature extractor computed on the same glass breaking event at different values of SNR (from $-5dB$ to $30dB$ in steps of $5dB$). The response keeps stable for positive, also very low values of SNR and it slightly decreases for null or negative SNR values. As demonstrated by the results that we reported in Section 5.4, the stability of the response of COPE feature extractors and the high performance on sounds with different values of SNR. The decrease of

performance at null and negative SNR is due to the effect of background sounds with energy higher than that of the sounds of interest. It determines strong changes of the position of the energy peaks with respect to those determined in the configuration. To this concern, the effect of other functions $\psi(f, t)$ in eq. 6 to evaluate the energy peak similarity can be explored.

7. Conclusions and outlook

We proposed a novel method for feature extraction in audio signals based on trainable feature extractors, which we called COPE (that stands for Combination of Peaks of energy). We employed the COPE feature extractors in the task of environmental sound event detection and classification, and tested their robustness to variations of the SNR of the sounds of interest. The results that we achieved on four public data sets (recognition rate equal to 91.71% on the MIVIA audio events, 94% on the MIVIA road events, 81.25% on the ESC-10 and 94.27% on the TU Dortmund data sets) are higher than many existing approaches and demonstrate the effectiveness of the proposed method.

The design of COPE feature extractors was based on neuro-physiological evidence of the mechanism that translates sound pressure waves into neural stimuli from the cochlea membrane through the Inner Hair Cells (IHC) in the auditory system of mammals. The proposed method can be extended by also including in its processing the implementation of a neuron response inhibition mechanism that prevents the short-time firing of those IHCs that have recently fired [78]. In this view, the computation of the energy peak map would need to account for the energy distribution of the sounds of interest in each frequency band at a larger time scale, instead of performing a local analysis only. The extension of the COPE feature extractor with such inhibition phenomenon can further improve the robustness of the proposed method to changes of background noise and SNR, as only significant energy peaks are to be processed.

Although the current implementation of the COPE feature extractor is rather efficient (0.965 seconds to compute a COPE feature vector of 200 ele-

ments for a signal of 3 seconds, on a 2GHz dual core CPU), their computation can be further speeded-up. Parallelization approaches can be explored, which compute the value of COPE features or the local energy peak responses in separate threads. The construction of the COPE feature vector can also be optimized by including in the classification system only those filters that are relevant for the application at hand. A feature selection scheme based on the relevance of the feature values described can be employed [79]. The optimization of the number of configured feature extractors and the implementation of parallelization strategies can jointly contribute to the implementation of a real-time system for intelligent audio surveillance on edge embedded systems.

Appendix A. Biological motivation

The sound pressure waves that hit our ears are directed to the *cochlea* membrane in the inner auditory system. Different parts of the cochlea membrane vibrate according to the energy of the frequency components of the sound pressure waves [54]. A bank of Gammatone filters was proposed as a model of the cochlea membrane, whose response over time forms a spectrogram-like image called Gammatonegram [37]. The membrane vibrations stimulate firing of *inner hair cells (IHC)*, which are neurons that lay behind the cochlea. The firing activity of IHCs stimulates various fibers of the auditory nerve over time. We consider the pattern of the IHC firing activity as a descriptor of the input sound.

Given a prototype sound, a COPE feature extractor models the pattern of points that describe the IHC firing activity. We consider the points of highest local energy in the Gammatonegram as the locations at which the IHCs fire, and the constellation that they form is a robust representation of the pattern of interest. Hence, a COPE feature extractor is configured by modeling the constellation of the peak points of the Gammatonegram of a prototype sound.

References

[1] M. Crocco, M. Cristani, A. Trucco, V. Murino, Audio surveillance: A systematic review, ACM Comput. Surv. 48 (4) (2016) 52:1–52:46. doi:10.1145/2871183.

[2] M. Vacher, F. Portet, A. Fleury, N. Noury, Challenges in the processing of audio channels for ambient assisted living, in: IEEE Healthcom, 2010, pp. 330–337.

[3] J. C. Wang, H. P. Lee, J. F. Wang, C. B. Lin, Robust environmental sound recognition for home automation, IEEE Trans. Autom. Sci. Eng 5 (1) (2008) 25–31.

[4] J. Maxime, X. Alameda-Pineda, L. Girin, R. Horaud, Sound representation and classification benchmark for domestic robots, in: IEEE ICRA, 2014, pp. 6285–6292. doi:10.1109/ICRA.2014.6907786.

[5] L. Besacier, E. Barnard, A. Karpov, T. Schultz, Automatic speech recognition for under-resourced languages: A survey, Speech Communication 56 (0) (2014) 85–100. doi:10.1016/j.specom.2013.07.008.

[6] A. Roy, M. Magimai-Doss, S. Marcel, A fast parts-based approach to speaker verification using boosted slice classifiers, IEEE Trans. Inf. Forensics Security 7 (1) (2012) 241–254. doi:10.1109/TIFS.2011.2166387.

[7] Z. Fu, G. Lu, K. M. Ting, D. Zhang, A survey of audio-based music classification and annotation, IEEE Trans. Multimedia 13 (2) (2011) 303–319.

[8] M. Cowling, R. Sitte, Comparison of techniques for environmental sound recognition, Pattern Recogn. Lett. 24 (15) (2003) 2895–2907.

[9] S. Chachada, C. C. J. Kuo, Environmental sound recognition: A survey, in: APSIPA, 2013, pp. 1–9. doi:10.1109/APSIPA.2013.6694338.

[10] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444. doi:10.1038/nature14539.

[11] Y. G. Jui-Ting Huang, Jinyu Li, An analysis of convolutional neural networks for speech recognition, in: ICASSP, 2015.

[12] A. van den Oord, S. Dieleman, B. Schrauwen, Deep content-based music recommendation, in: NIPS, 2013, pp. 2643–2651.

[13] H. Phan, L. Hertel, M. Maass, P. Koch, R. Mazur, A. Mertins, Improved audio scene classification based on label-tree embeddings and convolutional neural networks, IEEE Trans. Audio, Speech, Language Process. 25 (6) (2017) 1278–1290.

[14] N. Strisciuglio, M. Vento, N. Petkov, Bio-inspired filters for audio analysis, in: BrainComp 2015, Revised Selected Papers, 2016, pp. 101–115. doi:10.1007/978-3-319-50862-7_8.

[15] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, Reliable detection of audio events in highly noisy environments, Pattern Recogn. Lett. 65 (2015) 22 – 28. doi:10.1016/j.patrec.2015.06.026.

[16] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, Audio surveillance of roads: A system for detecting anomalous sounds, IEEE Trans. Intell. Transp. Syst. 17 (1) (2016) 279–288. doi:10.1109/TITS.2015.2470216.

[17] K. J. Piczak, Esc: Dataset for environmental sound classification, in: Proc ACM Int Conf Multimed, MM '15, 2015, pp. 1015–1018.

[18] A. Plinge, R. Grzeszick, G. A. Fink, A bag-of-features approach to acoustic event detection, in: IEEE ICASSP, 2014, pp. 3704–3708.

[19] H. Liu, J. Lu, J. Feng, J. Zhou, Group-aware deep feature learning for facial age estimation, Pattern Recognition 66 (2017) 82 – 94.

[20] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, Exploiting the deep learning paradigm for recognizing human actions, in: IEEE AVSS 2014, 2014, pp. 93–98. doi:10.1109/AVSS.2014.6918650.

[21] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognition 48 (10) (2015) 2993 – 3003.

[22] L. G. Hafemann, R. Sabourin, L. S. Oliveira, Learning features for offline handwritten signature verification using deep convolutional neural networks, Pattern Recognition 70 (2017) 163 – 176. doi:10.1016/j.patcog.2017.05.012.

[23] Y. Aytar, C. Vondrick, A. Torralba, Soundnet: Learning sound representations from unlabeled video, in: NIPS 2016, 2016.

[24] R. Rubinstein, M. Zibulevsky, M. Elad, Double sparsity: Learning sparse dictionaries for sparse signal approximation, IEEE Transactions on Signal Processing 58 (3) (2010) 1553–1564. doi:10.1109/TSP.2009.2036477.

[25] Y. Chen, J. Su, Sparse embedded dictionary learning on face recognition, Pattern Recognition 64 (2017) 51 – 59.

[26] K. Bunte, M. Biehl, M. F. Jonkman, N. Petkov, Learning effective color features for content based image retrieval in dermatology, Pattern Recognition 44 (9) (2011) 1892 – 1902. doi:10.1016/j.patcog.2010.10.024.

[27] N. Passalis, A. Tefas, Neural bag-of-features learning, Pattern Recognition 64 (2017) 277 – 294. doi:10.1016/j.patcog.2016.11.014.

[28] P. Koniusz, F. Yan, P. H. Gosselin, K. Mikolajczyk, Higher-order occurrence pooling for bags-of-words: Visual concept detection, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2017) 313–326. doi:10.1109/TPAMI.2016.2545667.

[29] G. Guo, S. Z. Li, Content-based audio classification and retrieval by support vector machines, IEEE Trans. Neur. Netw. 14 (1) (2003) 209–215.

[30] C. Nadeu, D. Macho, J. Hernando, Time and frequency filtering of filter-bank energies for robust HMM speech recognition, Speech Communication 34 (2001) 93 – 114. doi:10.1016/S0167-6393(00)00048-0.

[31] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, A. Serralheiro, Non-speech audio event detection, in: IEEE ICASSP, 2009, pp. 1973–1976.

[32] C. Clavel, T. Ehrette, G. Richard, Events detection for an audio-based surveillance system, in: ICME, 2005, pp. 1306 –1309. doi:10.1109/ICME.2005.1521669.

[33] P. K. Atrey, N. C. Maddage, M. S. Kankanhalli, Audio based event detection for multimedia surveillance, in: IEEE ICASSP, Vol. 5, 2006.

[34] A. Rabaoui, M. Davy, S. Rossignol, N. Ellouze, Using one-class svms and wavelets for audio surveillance, IEEE Trans. Inf. Forensics Security 3 (4) (2008) 763–775.

[35] S. Chu, S. Narayanan, C. C. J. Kuo, Environmental sound recognition with time-frequency audio features, IEEE Trans. Audio, Speech, Language Process. 17 (6) (2009) 1142–1158. doi:10.1109/TASL.2009.2017438.

[36] J. Dennis, H. D. Tran, E. S. Chng, Image feature representation of the subband power distribution for robust sound event classification, IEEE Trans. Audio, Speech, Language Process. 21 (2) (2013) 367–377.

[37] R. D. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang, M. Allerhand, Complex Sounds and auditory images, in: Auditory Physiology and Perception, 1992, pp. 429–443.

[38] R. F. Lyon, J. Ponte, G. Chechik, Sparse coding of auditory features for machine hearing in interference, in: IEEE ICASSP, 2011, pp. 5876–5879.

[39] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, Cascade classifiers trained on gammatonegrams for reliably detecting audio events, in: IEEE AVSS, 2014, pp. 50–55.

[40] J.-J. Aucouturier, B. Defreville, F. Pachet, The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music, J Acoust Soc Am 122 (2) (2007) 881–891.

[41] S. Pancoast, M. Akbacak, Bag-of-audio-words approach for multimedia event classification., in: INTERSPEECH, 2012, pp. 2105–2108.

[42] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, M. D. Plumbley, A database and challenge for acoustic scene classification and event detection, in: EUSIPCO, 2013, pp. 1–5.

[43] X. Lu, Y. Tsao, S. Matsuda, C. Hori, Sparse representation based on a bag of spectral exemplars for acoustic event detection, in: IEEE ICASSP, 2014, pp. 6255–6259. doi:10.1109/ICASSP.2014.6854807.

[44] R. Grzeszick, A. Plinge, G. Fink, Temporal acoustic words for online acoustic event detection, in: Pattern Recognition, Vol. 9358 of LNCS, 2015, pp. 142–153.

[45] M. Chin, J. Burred, Audio event detection based on layered symbolic sequence representations, in: IEEE ICASSP, 2012, pp. 1953–1956.

[46] A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, B. Raj, Audio event detection from acoustic unit occurrence patterns, in: IEEE ICASSP, 2012, pp. 489–492.

[47] H. Phan, L. Hertel, M. Maass, R. Mazur, A. Mertins, Audio phrases for audio event recognition, in: EUSIPCO, 2015.

[48] R. Grzeszick, A. Plinge, G. A. Fink, Bag-of-features methods for acoustic event detection and classification, IEEE Trans. Audio, Speech, Language Process. 25 (6) (2017) 1242–1252. doi:10.1109/TASLP.2017.2690574.

[49] K. J. Piczak, Environmental sound classification with convolutional neural networks, in: IEEE MLSP, 2015, pp. 1–6. doi:10.1109/MLSP.2015.7324337.

[50] J. S. Downie, A. F. Ehmann, M. Bay, M. C. Jones, The Music Information Retrieval Evaluation eXchange: Some Observations and Insights, 2010, pp. 93–115.

[51] J. Barker, E. Vincent, N. Ma, H. Christensen, P. Green, The PASCAL CHiME speech separation and recognition challenge, Computer Speech and Language 27 (3) (2013) 621 – 633. doi:10.1016/j.csl.2012.10.004.

[52] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, P. Soundararajan, The CLEAR 2006 Evaluation, 2007, pp. 1–44.

[53] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, M. D. Plumbley, Detection and classification of acoustic scenes and events, IEEE Trans. Multimedia 17 (10) (2015) 1733–1746. doi:10.1109/TMM.2015.2428998.

[54] R. D. Patterson, B. C. J. Moore, Auditory filters and excitation patterns as representations of frequency resolution, Frequency selectivity in hearing (1986) 123–177.

[55] B. R. Glasberg, B. C. Moore, Derivation of auditory filter shapes from notched-noise data, Hearing Research 47 (12) (1990) 103 – 138.

[56] A. L. chun Wang, T. F. B. F, An industrial-strength audio search algorithm, in: ISMIR, 2003.

[57] A. Palmer, I. Russell, Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells, Hearing Research 24 (1) (1986) 1 – 15. doi:10.1016/0378-5955(86)90002-X.

[58] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297. doi:10.1007/BF00994018.

[59] K. Morik, P. Brockhausen, T. Joachims, Combining statistical learning with a knowledge-based approach – a case study in intensive care monitoring, in: International Conference on Machine Learning (ICML), 1999, pp. 268–277.

[60] A. Saggese, N. Strisciuglio, M. Vento, N. Petkov, Time-frequency analysis for audio event detection in real scenarios, in: AVSS, 2016, pp. 438–443. doi:10.1109/AVSS.2016.7738082.

[61] Y. Aytar, C. Vondrick, A. Torralba, Soundnet: Learning sound representations from unlabeled video, in: NIPS, 2016, pp. 892–900.

[62] F. Colangelo, F. Battisti, M. Carli, A. Neri, F. Calabr, Enhancing audio surveillance with hierarchical recurrent neural networks, in: AVSS, 2017, pp. 1–6. doi:10.1109/AVSS.2017.8078496.

[63] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, M. Vento, Audio surveillance using a bag of aural words classifier, in: IEEE AVSS, 2013, pp. 81–86. doi:10.1109/AVSS.2013.6636620.

[64] L. Hertel, H. Phan, A. Mertins, Comparing time and frequency domain for audio event recognition using deep learning, in: IJCNN, 2016, pp. 3407–3411.

[65] F. Medhat, D. Chesmore, J. Robinson, Environmental Sound Recognition Using Masked Conditional Neural Networks, 2017, pp. 373–385.

[66] C. Nadeau, Y. Bengio, Inference for the generalization error, Machine Learning 52 (3) (2003) 239–281. doi:10.1023/A:1024068626366.

[67] G. Azzopardi, N. Petkov, A CORF computational model of a simple cell that relies on LGN input outperforms the Gabor function model, Biological Cybernetics 106 (3) (2012) 177–189. doi:{10.1007/s00422-012-0486-6}.

[68] G. Azzopardi, N. Petkov, Trainable COSFIRE filters for keypoint detection and pattern recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 490–503. doi:10.1109/TPAMI.2012.106.

[69] G. Azzopardi, N. Strisciuglio, M. Vento, N. Petkov, Trainable COSFIRE filters for vessel delineation with application to retinal images, Medical Image Analysis 19 (1) (2015) 46 – 57. doi:10.1016/j.media.2014.08.002.

[70] N. Strisciuglio, G. Azzopardi, M. Vento, N. Petkov, Unsupervised delineation of the vessel tree in retinal fundus images, in: Computational Vision and Medical Image Processing VIPIMAGE 2015, 2015, pp. 149–155.

[71] N. Strisciuglio, N. Petkov, Delineation of line patterns in images using b-cosfire filters, in: IWOBI, 2017, pp. 1–6. doi:10.1109/IWOBI.2017.7985538.

[72] N. Strisciuglio, G. Azzopardi, N. Petkov, Detection of curved lines with b-cosfire filters: A case study on crack delineation, in: CAIP, 2017, pp. 108–120. doi:10.1007/978-3-319-64689-3_9.

[73] N. Strisciuglio, G. Azzopardi, N. Petkov, Brain-inspired robust delineation operator, in: Computer Vision – ECCV 2018 Workshops, 2019, pp. 555–565.

[74] A. Saggese, N. Strisciuglio, M. Vento, N. Petkov, Learning skeleton representations for human action recognition, Pattern Recognition Lettersdoi:10.1016/j.patrec.2018.03.005.

[75] M. Newton, L. Smith, Biologically-inspired neural coding of sound onset for a musical sound classification task, in: IJCNN, 2011, pp. 1386–1393. doi:10.1109/IJCNN.2011.6033386.

[76] A. Neocleous, G. Azzopardi, C. Schizas, N. Petkov, Filter-based approach for ornamentation detection and recognition in singing folk music, in: CAIP, Vol. 9256 of LNCS, 2015, pp. 558–569. doi:10.1007/978-3-319-23192-1_47.

[77] P. Cano, E. Batlle, T. Kalker, J. Haitsma, A review of audio fingerprinting, Journal of VLSI signal processing systems for signal, image and video technology 41 (3) (2005) 271–284. doi:10.1007/s11265-005-4151-3.

[78] E. A. Lopez-Poveda, A. Eustaquio-Martín, A biophysical model of the inner hair cell: The contribution of potassium currents to peripheral auditory compression, Journal of the Association for Research in Otolaryngology 7 (3) (2006) 218–235. doi:10.1007/s10162-006-0037-8.

[79] N. Strisciuglio, G. Azzopardi, M. Vento, N. Petkov, Supervised vessel delineation in retinal fundus images with the automatic selection of B-COSFIRE filters, Mach. Vis. Appl. (2016) 1–13doi:10.1007/s00138-016-0781-7.