Ludeña-Choez, J., & Gallardo-Antolín, A. (2016). Acoustic event classification using spectral band selection and non-negative matrix factorization-based features. Expert Systems with Applications, 46, 77-86.

doi:https://doi.org/10.1016/j.eswa.2015.10.018

Acoustic Event Classification Using Spectral Band Selection and Non-Negative Matrix Factorization-Based Features

Jimmy Ludeña-Choez[a,b], Ascensión Gallardo-Antolín[a,*]

[a] *Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, Leganés 28911, Madrid, Spain*
[b] *Facultad de Ingeniería y Computación, Universidad Católica San Pablo, Campus Campiña Paisajista s/n Quinta Vivanco, Barrio de San Lázaro, Arequipa, Perú*

Abstract

Feature extraction methods for sound events have been traditionally based on parametric representations specifically developed for speech signals, such as the well-known Mel Frequency Cepstrum Coefficients (MFCC). However, the discrimination capabilities of these features for Acoustic Event Classification (AEC) tasks could be enhanced by taking into account the spectro-temporal structure of acoustic event signals. In this paper, a new front-end for AEC which incorporates this specific information is proposed. It consists of two different stages: short-time feature extraction and temporal feature integration. The first module aims at providing a better spectral representation of the different acoustic events on a frame-by-frame basis, by means of the automatic selection of the optimal set of frequency bands from which cepstral-like features are extracted. The second stage is designed for capturing the most relevant temporal information in the short-time features, through the application of Non-Negative Matrix Factorization (NMF) on their periodograms computed over long audio segments. The whole front-end has been evaluated in clean and noisy conditions. Experiments show that the removal of certain frequency bands (which are mainly located in the medium region of the spectrum for clean conditions

*Corresponding author
Email addresses: `jimmy@tsc.uc3m.es`; `jludenac@ucsp.edu.pe` (Jimmy Ludeña-Choez), `gallardo@tsc.uc3m.es` (Ascensión Gallardo-Antolín)

Preprint submitted to Expert Systems with Applications *October 16, 2015*

and in low frequencies for noisy environments) in the short-time feature computation process in conjunction with the NMF technique for temporal feature integration improves significantly the performance of a Support Vector Machine (SVM) based AEC system with respect to the use of conventional MFCCs.

Keywords: acoustic event classification, feature extraction, temporal feature integration, feature selection, mutual information, non-negative matrix factorization

---

1. Introduction

In recent years, the problem of automatically detecting and classifying acoustic non-speech events has attracted the attention of numerous researchers. Although speech is the most informative acoustic event, other kind of sounds (such as laughs, coughs, keyboard typing, etc.) can give relevant cues about the human presence and activity in a certain scenario (for example, in an office room). This information could be used in different applications, mainly in those with perceptually aware interfaces such as smart-rooms (Temko & Nadeu, 2006), automotive applications (Muller et al., 2008), mobile robots working in diverse environments (Chu et al., 2006) or surveillance systems (Principi et al., 2015).

Acoustic Event Classification (AEC) systems can be formulated as a machine learning problem consisting in two main stages: feature extraction (or front-end) and classification (or back-end). The first one obtains a parametric and compact representation of the audio signals more appropriate for classification. The purpose of the second one is to determine which Acoustic Event (AE) has been produced through a certain decision process. Several front-ends and classifiers have been proposed and compared in the literature for this task. Nevertheless, the high correlation between the performance of different classifiers suggests that the main problem is not the choice of the classification technique, but a design of a suitable feature extraction process for AEC (Kons & Toledo, 2013). This paper, precisely, focuses on this issue.

Many state-of-the art front-ends are composed of two modules: *short-time* *feature extraction*, in which acoustic coefficients are computed on a frame-by-frame basis (typically, the frame period used for speech/audio analysis is about 10-20 ms) from analysis windows of 20-40 ms, and *temporal feature integration* (Meng et al., 2007), in which features at larger time scales are extracted by combining somehow the short-time characteristics information over a longer time-frame composed of several consecutive frames. The resulting characteristics are often called *segmental features* (Zhang & Schuller, 2012; Ludeña-Choez & Gallardo-Antolín, 2013a, 2015). In this paper, two techniques which improve the performance of each of these modules by taking into account the specific spectro-temporal structure of acoustic events are presented. For short-time feature extraction, an automatic spectral band selection method is applied in order to emphasize the more relevant frequencies (and less redundant) of the acoustic events in the parameterization procedure, whereas for temporal feature integration, Non-Negative Matrix Factorization (NMF) (Lee & Seung, 1999) is used for obtaining a set of segmental features which better summarizes the temporal information contained in the frame-based acoustic characteristics.

This paper is organized as follows: Section 2 introduces related work on feature extraction of acoustic event signals. Section 3 describes the short-time feature extraction process based on spectral band selection. Section 4 presents the application of NMF for the design of the temporal feature integration module. Section 5 presents the experiments and results to end with some conclusions in Section 6.

2. Related work

In first works on acoustic event classification and detection, the parametric representations of audio signals used were strongly based on those previously developed for speech processing and related tasks, such as speech and speaker recognition. As these acoustic parameters are usually extracted on a frame-by-frame basis, they are commonly known as short-time features. Good examples are the conventional Mel-Frequency Cepstral Coefficients (MFCC) (Temko &

Nadeu, 2006; Zieger, 2008; Zhuang et al., 2010; Kwangyoun & Hanseok, 2011), log filter bank energies (Zhuang et al., 2010), Perceptual Linear Prediction (PLP) (Portelo et al., 2009), log-energy, spectral flux, entropy and zero-crossing rate (Temko & Nadeu, 2006; Perperis et al., 2011). The combination of some of these short-time features into high-dimensional acoustic vectors has also been studied, as well as the application of feature selection algorithms over these large pools of characteristics, in order to precisely reduce their dimensionality (Zhuang et al., 2008, 2010; Butko & Nadeu, 2010; Kiktova-Vozarikova et al., 2013).

Nevertheless, as pointed in (Zhuang et al., 2010), many of these conventional acoustic features are not necessarily the more appropriate for AEC tasks because most of them have been designed according to the spectral characteristics of speech which are quite different from the spectral structure of acoustic events. In addition, some types of acoustic events present a typical temporal structure (for example, the periodic pattern of phone rings) that should be somehow exploited in order to improve feature representation and discrimination capabilities. For these two reasons, recent research is being focused on finding a set of features that adequately represents the acoustic events.

To deal with the first problem, new acoustic parameters such as Power Normalised Cepstral Coefficients (PNCC) (Principi et al., 2015) and those derived from Gammatone (Plinge et al., 2014) or Gammachirp filter banks (Alam et al., 2014) have been proposed. Other works try to discover the hidden structure of the acoustic data by means of the application of Non-Negative Matrix Factorization (NMF) or K-Singular Value Decomposition (KSVD) on audio spectrograms (Choi et al., 2015). In an alternative approach (Ludeña-Choez & Gallardo-Antolín, 2013a), from the analysis of the AE spectral characteristics, it was concluded the importance of medium and high frequencies for discriminating between different acoustic events, yielding to the design of a new front-end based on the high pass filtering of the audio signals, which achieves good results in clean and noisy conditions (Ludeña-Choez & Gallardo-Antolín, 2015). Note that all these approaches can be seen as different modifications of the conventional mel-scaled auditory filter bank which is applied on the audio spectrograms in the short-time feature extraction process.

Following the idea that some frequency bands may be more useful for distinguishing between different sounds than others, in this paper, a modified mel-scaled filter bank is proposed in which only a selected set of spectral bands are considered in the computation of the short-time characteristics. In contrast to the already mentioned approaches, in this work, an automatic method is used to find this optimal set of frequency bands from which cepstral-like coefficients are derived, as explained in Section 3. In particular, several Feature Selection (FS) techniques based on Mutual Information (MI) measures have been evaluated and compared for this purpose. Note that, in comparison with previous works about FS for tasks related to acoustic events, in this paper it is not intended to use FS for dimensionality reduction but to provide a better spectral representation of the AEs through the selection of the more relevant and less redundant spectral bands.

In order to cope with the second problem, the idea of simultaneously performing temporal and spectral analysis to yield so-called spectro-temporal features has lately emerged, e.g. high-level features (also called audio banks) (Sandhan et al., 2014), spectrogram patch modeling using Restricted Boltzman Machines (RBM) (Espi et al., 2014) and 2D Gabor-based biologically inspired features (Schroder et al., 2015). As these methods are usually very computational demanding, temporal feature integration techniques, in which features at larger time scales are extracted by combining the short-time parameters contained in long audio segments, have become an interesting alternative. Among these techniques, the approach based on Filter Bank Coefficients (FC), which was initially proposed for general audio and music genre classification (McKinney & Breebaart, 2003; Arenas-García et al., 2006; Meng et al., 2007), has been experimented for AEC with promising results (Mejía-Navarrete et al., 2011). Its main advantage is that it allows to capture the dynamic structure in the short-time features. The idea behind FC is to summarize the periodogram of each short-time feature dimension by computing the power in several predefined frequency bands using a filter bank, which is usually the one proposed in (McKinney & Breebaart, 2003). However, as pointed in (Arenas-García et al., 2006), this fixed filter bank is not general enough since the relevance of the dynamics in the short-time features for classification can be expected to be task-dependent.

Based on this premise, in (Ludeña-Choez & Gallardo-Antolín, 2013b) a method based on Non-Negative Matrix Factorization (NMF) for the design of a filter bank for the computation of FC-based features more suitable for AEC has been proposed by the authors and successfully tested in clean conditions. In comparison with similar works (Arenas-García et al., 2006), the approach described in (Ludeña-Choez & Gallardo-Antolín, 2013b), which is described in Section 4, is unsupervised and general enough to be applied to any sound signals.

In summary, in view of the main limitations of the audio feature extraction methods existing in the literature, in this paper, a novel front-end for AEC tasks is proposed. The major contributions of this work are the following: the development of a new short-time parameterization based on the automatic selection of spectral bands which better reflects the spectral characteristics of audio events, its combination with a feature integration technique based on NMF which aims to improve the modeling of the temporal behaviour of short-time features; and the evaluation of the complete front-end in both, clean and noisy conditions.

Figure 1 represents the block diagram of the whole audio feature extraction process. As mentioned before, it can be observed that it consists of two main stages: short-time feature extraction and temporal feature integration. Next sections are devoted to the description of both modules.

3. Short-time Feature Extraction Based on Spectral Band Selection

In this section, the procedure of extraction of short-time acoustic characteristics from audio signals is presented. The main idea of this module is that not all the available spectral bands should be used in the feature extraction process, as only some of them provide suitable information for the acoustic event clas-

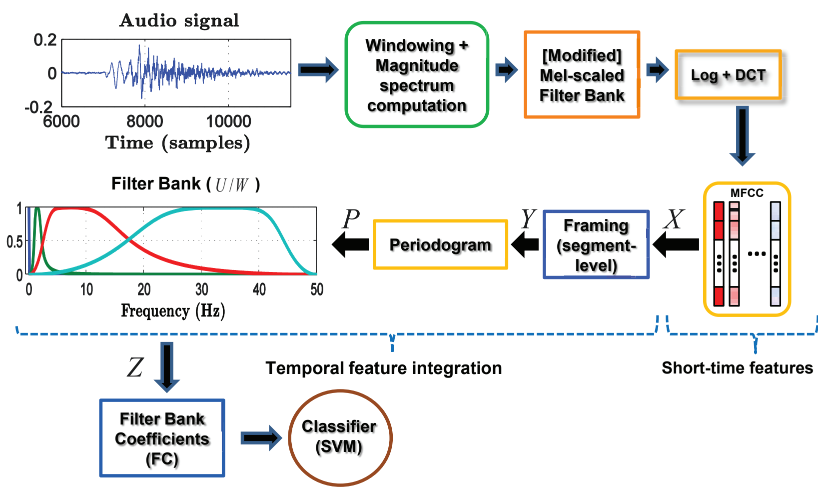

Figure 1: Block diagram of the feature extraction process.

sification task. As a consequence, in this approach, a method for choosing the most appropriate spectral bands is needed. In particular, in this work, several feature selection algorithms based on Mutual Information have been considered as it is explained in next subsection.

After a brief introduction about feature selection and its application to automatic spectral band selection, in the remainder of this section, the detailed process for obtaining a parametric representation of audio signals from the outputs of the selected frequency bands is described.

3.1. Feature Selection based on Mutual Information

The main objective of feature selection methods is to construct subsets of features that are useful for classification (Guyon & Elisseeff, 2003). They can be categorized as classifier-dependent (called "wrapper" methods) and classifier-independent (denoted as "filter" techniques) (Guyon & Elisseeff, 2003). Filter methods search the best feature sets by computing some similarity measures over the data, such as distance (Bins & Draper, 2001; Sebban & Nock, 2002) or mutual information (Peng et al., 2005; Férnandez et al., 2009; Brown et al., 2012), independently of any particular classifier, and therefore, they are less likely to overfit and less computationally costly than wrappers. For these reasons, in this paper, filter methods have been chosen, in particular those based on Mutual Information.

MI is a nature measure of the quantity of information that two random variables have in common. It is symmetric and non-negative and is zero if and only if the variables are independent (Cover & Thomas, 2006). MI can be seen as a way of quantify the relevance of one random variable with respect to the another one. Let $\mathbf{L}$ and $\mathbf{S}$ two discrete random variables and $\mathbf{l}$ and $\mathbf{s}$, two possible values adopted by, respectively, $\mathbf{L}$ and $\mathbf{S}$. The mutual information $\mathbf{I}(\mathbf{L};\mathbf{S})$ between $\mathbf{L}$ and $\mathbf{S}$ is given by

$$\mathbf{I}\left(\mathbf{L};\mathbf{S}\right) = \sum_{\mathbf{l}\in\mathbf{L}}\sum_{\mathbf{s}\in\mathbf{S}}\mathbf{p}\left(\mathbf{l},\mathbf{s}\right)\log\left(\frac{\mathbf{p}\left(\mathbf{l},\mathbf{s}\right)}{\mathbf{p}\left(\mathbf{l}\right)\mathbf{p}\left(\mathbf{s}\right)}\right) \tag{1}$$

where $\mathbf{p}\left(\mathbf{l}\right)$ and $\mathbf{p}\left(\mathbf{s}\right)$ are the probability distributions of $\mathbf{L}$ and $\mathbf{S}$ and $\mathbf{p}\left(\mathbf{l},\mathbf{s}\right)$ is their joint probability distribution.

FS methods based on MI rely on the definition of a certain selection criterion, $\mathbf{J}$, which is somehow related to the mutual information between features and classes and quantifies the usefulness of a feature subset for the classification task. Brown et al. (Brown et al., 2012) present an unifying view of several well-known MI-based FS techniques existing in the literature, showing that the criterion used in some of them can be expressed as linear combinations of MIs, as stated in (2),

$$\mathbf{J}\left(\mathbf{L_k}\right) = \mathbf{I}\left(\mathbf{L_k};\mathbf{S}\right) - \beta\sum_{\mathbf{L_j}\in\theta}\mathbf{I}\left(\mathbf{L_k};\mathbf{L_j}\right) + \gamma\sum_{\mathbf{L_j}\in\theta}\mathbf{I}\left(\mathbf{L_k};\mathbf{L_j}|\mathbf{S}\right) \tag{2}$$

where $\mathbf{L_k}$ is the feature to evaluate its inclusion in the feature set and θ is the set of currently selected features. The first term ensures the relevance of $\mathbf{L_k}$, the second term is related to the redundancy of $\mathbf{L_k}$ with features already selected in θ and the third term, called *conditional redundancy*, allows the inclusion of correlated features that, however, can be useful for the classification task. Different values of constants β and γ yield to different FS algorithms. In particular, in this work the following methods[1] have been considered:

- *Minimum-Redundancy Maximum-Relevance (mRMR)* ($\beta = \frac{1}{|\theta|}$, being $|\theta|$ the size of the current selected set, and $\gamma = 0$) which seeks to choose the features with highest relevance to the target class, whereas the redundancy is minimized (Peng et al., 2005).
- *Joint Mutual Information (JMI)* ($\beta = \frac{1}{|\theta|}$ and $\gamma = \frac{1}{|\theta|}$), which includes the conditional redundancy term to allow the inclusion of correlated features with complementary information (Meyer et al., 2008).
- *Conditional Informative Feature Extraction (CIFE)* ($\beta = 1$ and $\gamma = 1$), which also includes both, the redundancy and conditional redundancy terms, but with different weights than in JMI (Lin & Tang, 2006).
- *Conditional Redundancy (CondRed)* ($\beta = 0$ and $\gamma = 1$), which does not take into account the redundancy term.

3.2. Spectral Band Selection

In our case, for selecting the subset of spectral bands which better represents the different types of acoustic events, the input feature space for the FS algorithms consists of the log filter bank energies obtained after applying an auditory mel-scaled filter bank on the magnitude spectra of the instances of AEs belonging to the training partition of the database. In particular, these parameters are extracted every 10 ms using a Hamming analysis window of 20 ms long and a mel-scaled filter bank composed of 40 triangular bands which is the one implemented in the toolbox *VOICEBOX* (Brookes, 2009).

[1] Other criteria based on linear combinations of MIs, such as Mutual Information Feature Selection (MIFS) (Battiti, 1994) and non-linear combinations, such as Conditional Mutual Information Maximization (CMIM) (Fleuret, 2004) and Double Input Symmetrical Relevance (DISR) (Meyer & Bontempi, 2006), have also been tried. As these methods did not improve the results achieved by the ones described in this section and for the sake of brevity, they have not been included in the experimental section.

The four MI-based FS algorithms considered are applied over these data using the *FEAST* toolbox (Brown et al., 2012), in such a way that the variables involved in equations (1) and (2) are the mel-scaled log filter bank energies[2] $\mathbf{L} \in \mathbb{R}^N$ (being N the initial number of filters), and a discrete and finite set of acoustic event classes $\mathbf{S}$. After this process, for each FS method, a ranking of the selected spectral bands is obtained. As the chosen bands are finally sorted in ascending order, this mechanism can be seen as the modification of the original mel-scaled filter bank in which several filters are removed.

Note that the spectral band selection process is carried out only in the training stage of the system.

3.3. Short-Time Feature Computation

In the short-time feature extraction stage, audio signals are analyzed every 10 ms using a Hamming window of 20 ms long. For each window, the magnitude spectrum is obtained and filtered with the modified filter bank determined by the corresponding FS method, in such a way that only the log filter bank energies of the selected frequency bands are computed. Then, the resulting vector of log-energies is zero-padded to the number of filters of the original filter bank (in our case, 40) and a Discrete Cosine Transform (DCT) is applied over it, yielding to a set of 12 cepstral coefficients (C_1 to C_{12}). Note that in the case of using the complete mel-scaled filter bank (i.e. when none of the spectral bands is discarded), the resulting coefficients are the conventional MFCC. Finally, the log-energy of each frame (instead of the zero-order cepstrum coefficient) and the first time-derivatives are computed and added to the cepstral coefficients, leading to a 26-dimensional feature vector.

[2] As log filter bank energies are real values, an uniform quantization with 256 levels is performed over them, before the feature selection process itself.

4. NMF-Based Temporal Feature Integration

In this section, the background of the temporal feature integration technique called Filter bank Coefficients (FC) and its improvement by means of the use of Non-Negative Matrix Factorization are presented.

4.1. Filter bank Coefficients (FC)

Once the short-time acoustic characteristics are extracted, temporal feature integration is applied over audio segments of a given length (in our case, 2 s with overlap of 1 s) in order to obtain a set of feature vectors at a larger time scale (see Figure 1). In this work, the approach called Filter Bank Coefficients (FC) (McKinney & Breebaart, 2003; Arenas-García et al., 2006; Meng et al., 2007) is adopted, whose main advantage is that it aims at capturing the temporal short-time features' behaviour.

First, the sequence of T short-time coefficients of dimension D_x, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ is divided into K segments, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_K\}$ as follows,

$$\mathbf{y}_k = \{\mathbf{x}_{k \cdot H_s}, \mathbf{x}_{k \cdot H_s + 1}, ..., \mathbf{x}_{k \cdot H_s + L_s - 1}\} \tag{3}$$

where L_s is the segment size and H_s is the hop size, both defined in number of short-time frames.

Second, the periodogram of each dimension of the short-time features contained in the k-th segment $\mathbf{y}_k$ is estimated and, then, it is summarized by calculating the power in different frequency bands using a predefined filter bank,

$$\mathbf{z}_k = \mathbf{P}_k \mathbf{U} \tag{4}$$

where $\mathbf{P}_k$ comprises the periodograms of the sequence of the short-time coefficients belonging to the k-th segment, $\mathbf{U}$ is the frequency magnitude response of the FC filter bank and $\mathbf{z}_k$ is the final segmental feature vector. The dimensions of $\mathbf{P}_k$, $\mathbf{U}$ and $\mathbf{z}_k$ are, respectively, D_x x D_p, D_p x n_f and D_x x n_f, where D_p is the dimensionality of each individual periodogram and n_f is the number of filters in the bank. The FC parameters $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_K\}$ are the input to the AEC system, which, in this case, is based on Support Vector Machines (SVM).

Previous works (McKinney & Breebaart, 2003; Meng et al., 2007), in which the FC approach has been applied for general audio and music genre classification tasks, use a filter bank $\mathbf{U}$ composed of four filters corresponding to the following frequency bands:

- Filter 1: 0 Hz (DC filter)
- Filter 2: 1 - 2 Hz (modulation energy)
- Filter 3: 3 - 15 Hz (modulation energy)
- Filter 4: 20 - 43 Hz (perceptual roughness)

As the importance of the different dynamics in short-time features for classification may depend on the task, it can be argued that this fixed filter bank is not optimal for all audio classification problems. In other words, some modulation frequencies can be relevant for distinguishing between, for example, different acoustic events, and not between music genres. In next subsection, the unsupervised method developed by the authors for designing the FC filter bank is presented. More details about this method can be found in (Ludeña-Choez & Gallardo-Antolín, 2013b).

4.2. NMF-Based Design of the FC Filter Bank

For the improvement of the temporal feature integration module, the main goal is to develop an unsupervised approach to find the optimal filter bank in such a way that the resulting FC parameters $\mathbf{z}$ carry the most significant information about the underlying temporal structure of the short-time acoustic characteristics. This problem can be formulated as the decomposition of the periodograms $\mathbf{P}$ into their main components (i.e., into their more relevant frequency bands).

Non-Negative Matrix Factorization (NMF) (Lee & Seung, 1999) provides a way to decompose a signal into a convex combination of non-negative building blocks (called Spectral Basis Vectors, SBV) by minimizing a given cost function. As both, the power spectrum of the short-time parameters and the frequency response of the elements of the filter bank, are inherently positive, NMF can offer a suitable solution to the problem stated here, as will be explained in next subsections. Along the rest of the paper, the filter bank obtained by NMF is denoted as $\mathbf{W}$ in order to distinguish it from the fixed filter bank $\mathbf{U}$.

4.3. Non-Negative Matrix Factorization (NMF)

Given a matrix $\mathbf{V} \in \mathbb{R}_+^{A \times B}$, where each column is a data vector, NMF approximates it as a product of two matrices of non-negative low rank $\mathbf{W}$ and $\mathbf{H}$, such that

$$\mathbf{V} \approx \mathbf{WH} \tag{5}$$

where $\mathbf{W} \in \mathbb{R}_+^{A \times C}$ and $\mathbf{H} \in \mathbb{R}_+^{C \times B}$ and normally $C \leq min\,(A, B)$. This way, each column of $\mathbf{V}$ can be written as a linear combination of the C basis vectors (columns of $\mathbf{W}$), weighted with the coefficients of activation or gains located in the corresponding column of $\mathbf{H}$. NMF can be seen as a dimensionality reduction of data vectors from an $A-$dimensional space to a $C-$dimensional space. This is possible if the columns of $\mathbf{W}$ uncover the latent structure in the data (Lee & Seung, 1999). The factorization is achieved by an iterative minimization of a given cost function as, for example, the Euclidean distance or the generalized Kullbak Leibler (KL) divergence which is defined as follows,

$$D_{\mathrm{KL}}\left(\mathbf{V} \| \mathbf{WH}\right) = \sum_{ij} \left(\mathbf{V}_{ij} log \frac{\mathbf{V}_{ij}}{\left(\mathbf{WH}\right)_{ij}} - \left(\mathbf{V} - \mathbf{WH}\right)_{ij} \right) \tag{6}$$

In this work, the KL divergence is considered because it has been recently used with good results in speech processing tasks, such as speech enhancement and denoising for ASR tasks (Wilson et al., 2008; Ludeña-Choez & Gallardo-Antolín, 2012) or feature extraction (Schuller et al., 2010). In order to find a local optimum value for the KL divergence between $\mathbf{V}$ and $(\mathbf{WH})$, an iterative scheme with multiplicative update rules can be used as proposed in (Lee & Seung, 1999) and stated in (7),

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\frac{\mathbf{V}}{\mathbf{WH}}\mathbf{H}^T}{\mathbf{1H}^T} \qquad \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T\frac{\mathbf{V}}{\mathbf{WH}}}{\mathbf{W}^T\mathbf{1}} \tag{7}$$

where $\mathbf{1}$ is a matrix of size $\mathbf{V}$, whose elements are all ones and the multiplications $\otimes$ and divisions are component wise operations. NMF produces a sparse representation of the data, reducing the redundancy.

4.4. Constructing the FC Filter Bank with NMF

As mentioned before, the matrix to be decomposed is formed by the periodograms of the short-time acoustic characteristics. As a unique filter is learnt for all their components, the matrix $\mathbf{P}$ consists of the row-wise concatenation of the D_x periodograms of the short-time parameters extracted from the training set of the different acoustic events considered. Therefore, the dimension of $\mathbf{P}$ is $(D_x \text{ x } n_s)$ x D_p, where n_s is the total number of segments in the training set.

Once this matrix is transposed ($\mathbf{P^T}$), its corresponding factored matrices $\mathbf{WH}$ are obtained using the learning rules in equation (7). The dimensions of $\mathbf{W}$ and $\mathbf{H}$ are, respectively, D_p x n_f and n_f x $(D_x \text{ x } n_s)$. The resulting matrix $\mathbf{W}$ contains the SBVs which represent the basis of the power spectrum of the short-time features, as it is verified that $\mathbf{P^T} \approx \mathbf{WH}$, and, therefore, they could be interpreted as the filters of the required FC filter bank.

In order to compute the NMF-based FC parameters, equation (4) is applied substituting the fixed filter bank $\mathbf{U}$ by $\mathbf{W}$.

5. Experiments and Results

5.1. Database and Baseline System

The database used for the experiments consists of a total of 2,114 instances of target events belonging to 12 different acoustic classes: *Applause*, *Cough*, *Chair moving*, *Door knock*, *Door open/slam*, *Keyboard typing*, *Laugh*, *Paper* *work*, *Phone ring*, *Steps*, *Spoon/cup jingle* and *Key jingle*. The composition of the whole database was intended to be similar to the one used in (Zhuang et al., 2010) and it is shown in Table 1. Audio files were obtained from different sources: websites, the FBK-Irst database (FBK-Irst, 2009) and the UPC-TALP database (UPC-TALP, 2012). All sounds were converted to the same format and sampling frequency (8 KHz).

Table 1: *Database used in the experiments.*

Class	Event type	No. of occurrences
1	Applause [ap]	155
2	Cough [co]	199
3	Chair moving [cm]	115
4	Door knock [kn]	174
5	Door open/slam [ds]	251
6	Keyboard typing [kt]	158
7	Laugh [la]	224
8	Paper work [pw]	264
9	Phone ring [pr]	182
10	Steps [st]	153
11	Spoon/cup jingle [cl]	108
12	Key jingle [kj]	131
Total		2,114

Since this database is too small to achieve reliable classification results, a 6-fold cross validation was used in order to artificially extend it, averaging the results afterwards. Specifically, the database was split into six disjoint balanced groups, in such a way that one different group was kept for testing in each fold, while the remainder ones were used for training.

For the experiments in noisy conditions, the original audio recordings were contaminated with six different types of noise (*Airport*, *Babble*, *Restaurant*, *Train*, *Exhibition Hall* and *Subway*) obtained from the AURORA framework

(Pearce & Hirsch, 2000) at SNRs from 0 dB to 20 dB with 5 dB step. In order to calculate the amount of noise to be added to the clean recordings, the audio and noise powers were calculated following the procedure indicated in (Steeneken, 1991), which takes into account the non-stationary characteristics of the signals.

The AEC system is based on a one-against-one SVM with Radial Basis Function (RBF) kernel on normalized features (Ludeña-Choez & Gallardo-Antolín, 2013b, 2015). The system was developed using the LIBSVM software (Chang & Lin, 2011). Concerning SVM training, for each one of the subexperiments, a 5-fold cross validation was used for computing the optimal values of the RBF kernel parameters. In the testing stage, as the SVM classifier was fed with segmental features computed over sliding windows, the classification decisions were made at segment level. In order to obtain a decision for the whole instance (target event level), the classifier outputs of the corresponding windows were integrated using a majority voting scheme, in such a way that the most frequent label was finally assigned to the whole recording (Geiger et al., 2013).

5.2. Application of FS to Spectral Band Selection

For each fold, the selection of the more appropriate frequency bands for AEC was performed following the procedure described in Subsection 3.2. Cells in blue color in Figure 2 represent the 12 first non-selected spectral bands determined by mRMR, JMI, CIFE and CondRed algorithms for the first fold. The number inside each cell indicates the position in the rank of the discarded bands (for example, the 30^{th} band is the first discarded one by the mRMR algorithm). The non-selected bands do not differ very much between folds.

From this figure, it can be observed the following behaviour of the FS methods. CondRed discards the first low-frequency filters (this is equivalent to the high pass filtering approach proposed in (Ludeña-Choez & Gallardo-Antolín, 2015)). The non-selected bands by JMI are placed into two different frequency regions, the first one from 530 Hz to 1530 Hz and the second one from 2125 Hz to 2685 Hz. mRMR discards several non-adjacent bands in the spectral region

CondRed	1	2	3	4	5	6	7	8	9	10	11	12								
CIFE												11		9		7			6	
JMI													10	5	1	3	8		4	
mRMR																11			8	
# Band	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

CondRed																				
CIFE	2	1	4	10	5		3	8	12											
JMI	2	7		11							9	6	12							
mRMR	6		4	10	2			7		1		3	9		5	12				
# Band	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

Figure 2: Spectral bands discarded by different MI-based feature selection methods for the first fold training set.

between 920 Hz to 3200 Hz. Finally, CIFE does not select bands in an almost continuous region between 920 Hz to 2110 Hz and more sparsely between 650 Hz to 825 Hz.

5.3. Application of NMF to the Design of the FC Filter Bank

The filters of the fixed filter bank **U** were implemented as 2$^{\text{nd}}$ order Butterworth filters. On the contrary, in the NMF-based method, for each fold, the filter bank **W** was obtained by applying the method described in Subsection 4.4 over the corresponding training set. In all folds, NMF was initialized by generating 10 random matrices (**W** and **H**), in such a way that the factorization with the smallest euclidean distance between $\mathbf{P}^{\mathbf{T}}$ and (**W H**) was chosen for initialization. Then, these initial matrices were refined by minimizing the KL divergence using the multiplicative update rules given in equation (7) and a maximum of 200 iterations. After this process, the resulting **W** contained the filters of the required FC bank.

Figure 3 (b) represents the NMF-based FC filter bank **W** obtained on a single fold using the previous procedure for $n_f = 4$ filters. For comparison purposes, the fixed FC filter bank **U** is also represented in Figure 3 (a). Note that, although the maximum modulation frequency is 50 Hz (the short-time features are extracted each 10 ms), for improving the readability of the figures, only frequencies up to 20 Hz are represented. From the comparison of Figures 3 (a) and (b), it can be seen that filters 1 and 2 of **U** roughly appears in **W**.

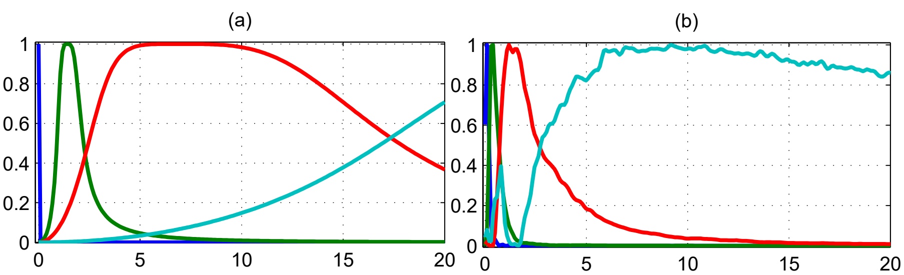

Figure 3: Frequency responses of the FC filter banks used in the temporal feature integration process. (a) Fixed filter bank (**U**); (b) Filter bank determined by NMF (**W**).

The highest frequency filter in **W** presents a high bandwidth and covers the modulation frequencies of the baseline filters 3 and 4. Finally, the filter 4 of **U** is substituted by a low-frequency filter in **W**, suggesting that, for describing the temporal structure of the short-time acoustic characteristics, low modulation frequencies are more relevant than high ones. Also, it is worth mentioning that the resulting filters do not differ very much between folds.

5.4. Results in Clean Conditions

This section contains the experiments carried out in order to assess the performance of the proposed front-end in clean conditions (when no noise is added to the audio signals) in comparison to the case in which the complete mel-scaled filter bank is used. For temporal feature integration, two different techniques have been evaluated, FC (with the fixed FC filter bank **U**) and FC_NMF (with the NMF-based FC filter bank **W**). The term "baseline" refers to the case in which the short-time features correspond to the conventional MFCC (i.e. when the complete mel-scaled filter bank is used in the short-time feature extraction stage). Therefore, the baseline for FC is the combination of MFCC for short-time feature extraction and FC for temporal feature integration. In the same way, for FC_NMF, the baseline is the combination of MFCC and FC_NMF.

The average Recognition Rate (RR), i.e., the percentage of target events correctly classified, of the baseline systems is 71.75% for FC and 73.15% for

FC_NMF. Figures 4 (a) and (b) represent, respectively, the Relative Error Reductions (RERs) with respect to the corresponding baselines for the FC and FC_NMF front-ends as a function of the number of discarded bands by the four FS algorithms considered: mRMR, JMI, CIFE and CondRed.

As it can be observed, for the FC parameterization, to consider only the most important spectral bands for the computation of the short-time features always outperforms the baseline, specially when the number of non-selected bands is in the range between 6 and 12. With respect to the performance of the different FS techniques, CondRed produces smaller improvements than the remaining algorithms, whereas mRMR and JMI achieve more similar results. CIFE is the method which produces the best performance with RERs with respect to the baseline between 16% and 19% when more than 5 bands are discarded.

In general terms, FC_NMF follows similar trends than FC, although the relative error reductions are more noticeable. Again, the smallest improvements are obtained with CondRed. However, in this case, JMI produces the best results, achieving RERs over 26% in the range from 7 to 9 non-selected spectral bands. Anyway, in both front-ends, it seems that the FS algorithms which exhibit better performance are those in which the redundancy and conditional redundancy terms are taken into account (JMI and CIFE). In these cases, the frequency bands not considered in the short-time feature extraction process are mainly in the medium region of the spectrum.

Table 2 shows the average recognition rates, as well as the corresponding 95% confidence intervals, achieved by FC and FC_NMF, for the baselines and the best configuration of the different FS methods. For both feature temporal integration techniques, spectral band selection improves significantly the baseline systems. For the FC front-end, CIFE with 12 discarded spectral bands obtains the best results, whereas for FC_NMF, the highest classification rate corresponds to JMI and 7 non-selected bands. In both cases, the improvement over the respective baselines is similar (around 5% absolute). Finally, comparing the accuracies with the best configurations, it can be observed that FC_NMF outperforms FC, being the performance differences statistically significant.

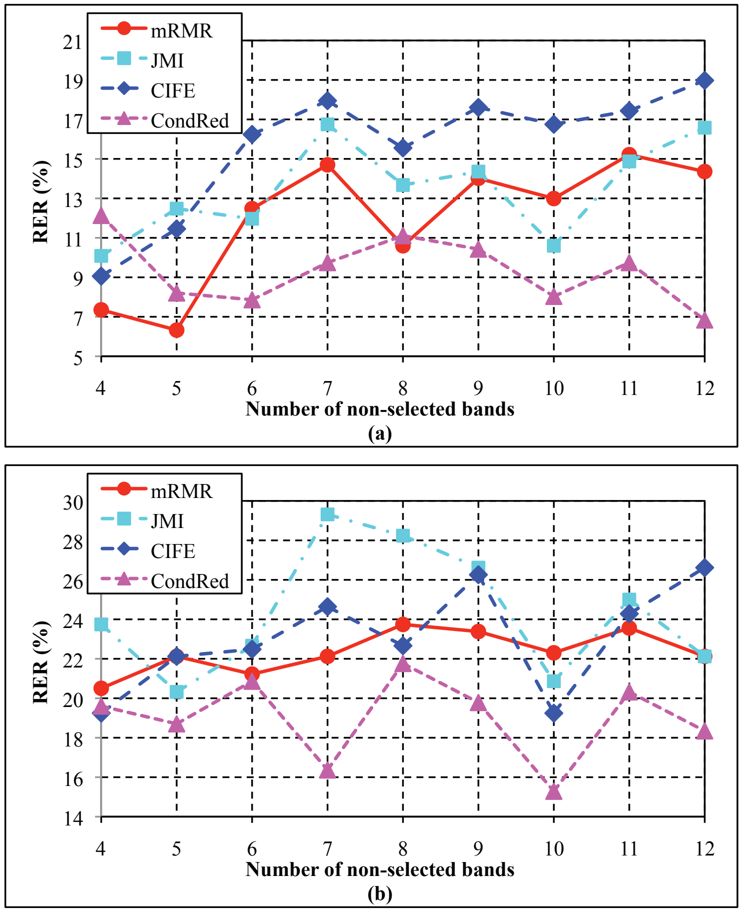

Figure 4: Relative error reduction [%] with respect to the corresponding baselines: (a) FC parameterization; (b) FC_NMF parameterization.

5.5. Results in Noisy Conditions

In order to study the impact of noisy environments on the performance of the AEC system, several experiments were carried out with six different types of noise (*Airport*, *Babble*, *Restaurant*, *Train*, *Exhibition Hall* and *Subway*) at SNRs from 0 dB to 20 dB with 5 dB step.

Table 3 shows the average recognition rates over all noises and SNRs considered, as well as the corresponding 95% confidence intervals, achieved by FC and FC_NMF, for the baseline and the best configuration of the different FS

Table 2: *Average recognition rates [%] for different FS methods and the FC and FC_NMF parameterizations in clean conditions.*

Short-time Features	Temporal Feature Integration			
	FC		FC_NMF	
	Average RR [%]	No. discarded bands	Average RR [%]	No. discarded bands
Baseline	71.75 ± 1.92	-	73.15 ± 1.89	-
mRMR	76.05 ± 1.82	11	79.53 ± 1.72	8
JMI	76.48 ± 1.81	7	$\mathbf{81.02 \pm 1.67}$	7
CIFE	$\mathbf{77.11 \pm 1.79}$	12	80.30 ± 1.70	12
CondRed	75.18 ± 1.84	4	79.00 ± 1.74	8

methods. A general comment is that in noisy conditions, a dramatic decrease in the classification rates is produced. As in clean conditions, for both temporal feature integration techniques, spectral band selection improves significantly the respective baseline systems. However, in this case, whereas mRMR, JMI and CIFE achieve similar recognition rates, CondRed produces better results than the remainder FS methods, being the performance differences statistically significant. In particular, the relative error reduction of CondRed with respect to the respective baselines is around 13% for FC when 5 bands are discarded and around 16% for FC_NMF for 9 discarded bands. Note that CondRed does not take into account the first low frequency filters of the auditory filter bank in the short-time feature extraction process. As in the selection process CondRed does not penalize features which are redundant with the other ones already chosen ($\beta = 0$), it seems that keeping spectral bands carrying similar information in clean conditions, can increase the robustness to noise of the whole system. This is because, when a certain frequency band is masked by the presence of noise, its spectral information is not completely lost if another redundant band has been preserved in the parameterization process.

When comparing the results obtained by FC and FC_NMF with the best configurations, it can be observed that FC_NMF improves the average recognition rates achieved by FC, being the performance differences statistically significant.

Figure 5 represents the recognition rates achieved by the baseline and the best configurations of the four FS methods with the FC_NMF front-end as a function of the SNR for the six noises evaluated. It can be observed that, in general, FS methods outperform the baseline for all noises and SNRs. mRMR, JMI and CIFE obtain similar results, whereas the classification rates achieved by CondRed are noticeably higher than those produced by the remaining FS methods and the baseline for the *Airport*, *Babble*, *Restaurant* and *Train* noises. For *Exhibition Hall* and *Subway* noises, CondRed still obtains the best results, but in these cases, the differences are smaller.

Table 3: *Average recognition rates [%] for different FS methods and the FC and FC_NMF parameterizations in noisy conditions.*

	Temporal Feature Integration			
	FC		FC_NMF	
Short-time Features	Average RR [%]	No. discarded bands	Average RR [%]	No. discarded bands
Baseline	45.54 ± 0.39	-	44.85 ± 0.39	-
mRMR	50.87 ± 0.39	6	51.92 ± 0.39	3
JMI	50.88 ± 0.39	5	51.20 ± 0.39	6
CIFE	51.13 ± 0.39	6	51.37 ± 0.39	5
CondRed	$\mathbf{52.41 \pm 0.39}$	5	$\mathbf{53.43 \pm 0.39}$	9

6. Conclusions

In this paper, a new front-end for acoustic event classification whose design incorporates information about the specific spectro-temporal patterns of acoustic events is proposed. It presents a modular structure consisting of two different stages: short-time feature extraction and temporal feature integration.

The first module is based on the selection of the optimal set of frequency bands which provides a better spectral representation of the different acoustic events and improves its discrimination capabilities compared to conventional MFCCs. This procedure is accomplished by means of the use of mutual information-based feature selection algorithms (mRMR, JMI, CIFE and CondRed) over the mel-scaled log filter bank energies. Once the log filter bank

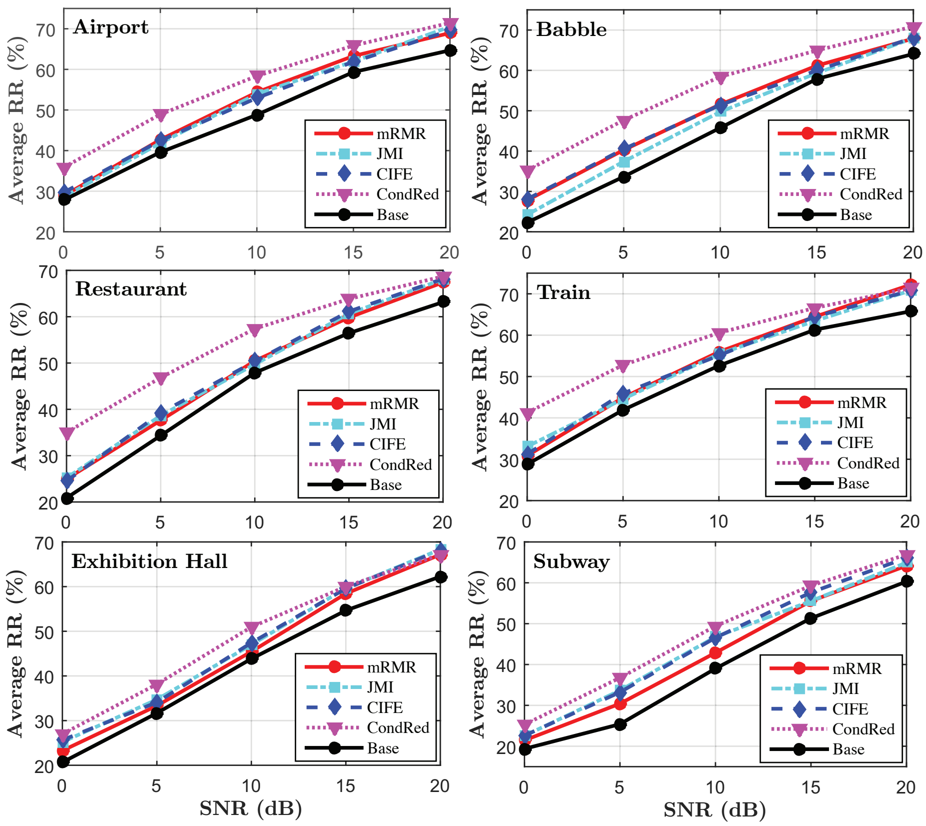

Figure 5: Average recognition rates for different noises and the FC_NMF parameterization.

energies of the chosen filters are extracted, the DCT is applied over them, yielding to a set of short-time cepstral-like coefficients, which are finally combined at a larger temporal scale through a process of temporal feature integration which is performed in the second module of the front-end. This stage relies on the combination of the feature integration technique called FC and non-negative matrix factorization, producing a set of segmental features called FC_NMF. In particular, NMF is used for the unsupervised learning of the filter bank which allows a better modeling of the temporal dynamics of the short-time parameters, in such a way that more reliable information about the temporal structure of the acoustic events is incorporated in the feature extraction process in comparison to the baseline FC technique.

The whole front-end has been tested in clean and noisy conditions on a SVM-based AEC system. On the one hand, the FS methods which achieve the best performance are CIFE and JMI for, respectively, the FC and FC_NMF parameterizations in clean conditions and CondRed in noisy conditions. Any way, it is shown that the removal of the frequency bands determined by the FS algorithms (which are mainly located in the medium region of the spectrum for clean conditions and in low frequencies for noisy environments) in the short-time feature computation process, improves significantly the performance of the baseline system (when no spectral bands are removed). On the other hand, the combination of these short-time acoustic characteristics with the FC_NMF technique produces significant improvements in the classification performance of the whole system in comparison with the FC-based features. This result suggests that NMF is able to better model the temporal behaviour of the short-time features than the conventional FC technique and that low modulation frequencies are more important than the high ones for distinguishing between different acoustic events.

As mentioned before, the central idea behind the proposed front-end is to take advantage of the specific spectral and temporal patterns of acoustic events for enhancing the representation and discrimination capabilities of the extracted features. Compared to previous related work in which a simultaneous spectro-temporal processing is performed (Espi et al., 2014), (Schroder et al., 2015), the main advantages of our system is that it is modular, so it is possible to independently optimized each stage and less computationally costly. Regarding the short-time feature extraction, in contrast to previous approaches in which the optimum set of spectral bands was manually determined(Ludeña-Choez & Gallardo-Antolín, 2015), our method automatically selects the most relevant bands and derives from them a set of decorrelated cepstral-like coefficients instead of directly using the log filter bank energies (Kiktova-Vozarikova et al., 2013). With respect to the temporal feature integration stage, the proposed technique models in a more adequate way the temporal dynamics of short-time features, as the filter bank used for this purpose is automatically learnt from data, in opposition to previous works in which this filter bank is fixed (Meng et al., 2007) and not necessarily adapted to the characteristics of the audio signals to be processed. In addition, results have shown that our system outperforms the baseline in both, clean and noisy scenarios, whereas many of previous related works have been tested only in clean conditions (for example, (Plinge et al., 2014; Sandhan et al., 2014)).

One of the disadvantages of the proposed front-end is that in its design, some interesting properties of the human auditory system, such as temporal and frequency masking, have not been taken into account. Nevertheless, for future work, this problem could be (at least, partially) overcome through the use of morphological operations on the spectrograms (de-la Calle-Silos et al., 2015) in the first stage of the parameterization scheme. Another limitation is that in the second module of the front-end, a unique NMF-based filter bank is learnt and used for all the components of short-time features, which might not be a realistic assumption. For this reason, the design of one different NMF-filter bank for each short-time feature dimension will be further studied.

Acknowledgements

This work has been partially supported by the Spanish Government grant TEC2014-53390-P. Financial support from the Fundación Carolina and Universidad Católica San Pablo, Arequipa (Jimmy Ludeña-Choez) is thankfully acknowledged.

References

Alam, M. J., Kenny, P., & O'Shaughnessy, D. (2014). Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique. *Digital Signal Processing*, *29*, 147–157. doi:10.1016/j.dsp.2014.03.001.

Arenas-García, J., Larsen, J., Hansen, L. K., & Meng, A. (2006). Optimal filtering of dynamics in short-time features for music organization. In *Proceedings* *of the International Society for Music Information Retrieval Conference (ISMIR)* (pp. 290–295).

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, *5(4)*, 537–550. doi:10.1109/72.298224.

Bins, J., & Draper, B. (2001). Feature selection from huge feature sets. In *IEEE Proceedings on Computer Vision (ICCV)* (pp. 159–165). volume 2. doi:10.1109/ICCV.2001.937619.

Brookes, M. (2009). Voicebox: Speech processing toolbox for MATLAB. Website. URL: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. Accessed: October 2015.

Brown, G., Pocock, A., Zhao, M., & Luján, M. (2012). Conditional likehood maximisation: A unifying framework for information theoretic feature selection. *Machine Learning Research*, *13*, 27–66.

Butko, T., & Nadeu, C. (2010). On enhancing acoustic event detection by using feature selection and audiovisual feature-level fusion. In *Proceedings of the Workshop on Database and Expert Systems Applications (DEXA)* (pp. 271–275). doi:10.1109/DEXA.2010.61.

de-la Calle-Silos, F., Valverde, F., Gallardo-Antolín, A., & Peláez, C. (2015). Morphologically filtered power-normalized cochleograms as robust, biologically inspired features for ASR. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*, 2070–2080. doi:10.1109/TASLP.2015.2464691.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 27:1–27. URL: https://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed: October 2015.

Choi, W., Park, S., Han, D. K., & Ko, H. (2015). Acoustic event recognition using dominant spectral basis vectors. In *Proceedings of the International Conference of the Speech Communication Association (INTERSPEECH)* (pp. 2002–2006).

Chu, S., Narayanan, S., Kuo, C.-C. J., & Mataric, M. J. (2006). Where am I? Scene recognition for mobile robots using audio features. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)* (pp. 885–888). doi:10.1109/ICME.2006.262661.

Cover, T., & Thomas, J. (2006). *Elements of Information Theory, 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Espi, M., Fujimoto, M., Kubo, Y., & Nakatani, T. (2014). Spectrogram patch based acoustic event detection and classification in speech overlapping conditions. In *Proceedings of the 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)* (pp. 117–121). doi:10.1109/HSCMA.2014.6843263.

FBK-Irst (2009). FBK-Irst database of isolated meeting-room acoustic events. ELRA Catalog no. S0296.

Férnandez, R., Bonastre, J., Matrouf, D., & Calvo, J. (2009). Feature selection based on information theory for speaker verification. In *Lecture Notes in Computer Science (LNCS), Springer* (pp. 305–312). volume 5856. doi:10.1007/978-3-642-10268-4_36.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, *5*, 1531–1555.

Geiger, J. T., Schuller, B., & Rigoll, G. (2013). Large-scale audio feature extraction and SVM for acoustic scene classification. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 1–4). doi:10.1109/WASPAA.2013.6701857.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Kiktova-Vozarikova, E., Juhar, J., & Cizmar, A. (2013). Feature selection for acoustic events detection. *Multimedia Tools and Applications*, (pp. 1–21). doi:10.1007/s11042-013-1529-2.

Kons, Z., & Toledo, O. (2013). Audio event classification using deep neural networks. In *Proceedings of the International Conference of the Speech Communication Association (INTERSPEECH)* (pp. 1482–1486).

Kwangyoun, K., & Hanseok, K. (2011). Hierarchical approach for abnormal acoustic event classification in an elevator. In *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)* (pp. 89–94). doi:10.1109/AVSS.2011.6027300.

Lee, D., & Seung, H. (1999). Algorithms for non-negative matrix factorization. *Nature*, *401*, 788–791.

Lin, D., & Tang, X. (2006). Conditional infomax learning an integrated framework for feature extraction and fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 68–82). doi:10.1007/11744023_6.

Ludeña-Choez, J., & Gallardo-Antolín, A. (2012). Speech denoising using non-negative matrix factorization with kullback-leibler divergence and sparseness constraints. In *Advances in Speech and Language Technologies for Iberian Languages, Communications in Computer and Information Science (CCIS), Springer* (pp. 207–216). volume 328. doi:10.1007/978-3-642-35292-8_22.

Ludeña-Choez, J., & Gallardo-Antolín, A. (2013a). NMF-based spectral analysis for acoustic event classification tasks. In *Proceedings of the Workshop on Linear Speech Processing (NOLISP), Lecture Notes in Computer Science (LNAI), Springer* (pp. 9–16). volume 7911. doi:10.1007/978-3-642-38847-7_2.

Ludeña-Choez, J., & Gallardo-Antolín, A. (2013b). NMF-based temporal feature integration for acoustic event classification. In *Proceedings of the Con-*

ference of the International Speech Communication Association (INTERSPEECH) (pp. 2924–2928).

Ludeña-Choez, J., & Gallardo-Antolín, A. (2015). Feature extraction based on the high-pass filtering of audio signals for acoustic event classification. *Computer Speech and Language*, *30*, 32–42. doi:10.1016/j.csl.2014.04.001.

McKinney, M., & Breebaart, J. (2003). Features for audio and music classification. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* (pp. 151–158).

Mejía-Navarrete, D., Gallardo-Antolín, A., Peláez, C., & Valverde, F. (2011). Feature extraction assessment for an acoustic-event classification task using the entropy triangle. In *Proceedings of the International Conference of the Speech Communication Association (INTERSPEECH)* (pp. 309–312).

Meng, A., Ahrendt, P., & Larsen, J. (2007). Temporal feature integration for music genre classification. *Proceedings of the IEEE Transactions on Audio, Speech, and Language Processing*, *15*, 1654–1664. doi:10.1109/TASL.2007.899293.

Meyer, P., & Bontempi, G. (2006). On the use of variable complementarity for feature selection in cancer classification. In *Evolutionary Computation and Machine Learning in Bioinformatics. Lecture Notes in Computer Science (LNCS)* (pp. 91–102). volume 3907. doi:10.1007/11732242_9.

Meyer, P., Schretter, C., & Bontempi, G. (2008). Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, *2*, 261–274. doi:10.1109/JSTSP.2008.923858.

Muller, C., Biel, J., Kim, E., & Rosario, D. (2008). Speech-overlapped acoustic event detection for automotive applications. In *Proceedings of the In-*

ternational Conference of the Speech Communication Association (INTERSPEECH) (pp. 2590–2593).

Pearce, D., & Hirsch, H. G. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP/INTERSPEECH)* (pp. 29–32).

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 1226–1238. doi:10.1109/TPAMI.2005.159.

Perperis, T., Giannakopoulos, T., Makris, A., Kosmopoulos, D. I., Tsekeridou, S., Perantonis, S. J., & Theodoridis, S. (2011). Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies. *Expert Systems with Applications*, *38*, 14102–14116. doi:10.1016/j.eswa.2011.04.219.

Plinge, A., Grzeszick, R., & Fink, G. A. (2014). A bag-of-features approach to acoustic event detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 3704–3708). doi:10.1109/ICASSP.2014.6854293.

Portelo, J., Bugalho, M., Trancoso, I., Neto, J., Abad, A., & Serralheiro, A. (2009). Non speech audio event detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 1973–1976). doi:10.1109/ICASSP.2009.4959998.

Principi, E., Squartini, S., Bonfigli, R., Ferroni, G., & Piazza, F. (2015). An integrated system for voice command recognition and emergency detection based on audio signals. *Expert Systems with Applications*, *42*, 5668–5683. doi:10.1016/j.eswa.2015.02.036.

Sandhan, T., Sonowal, S., & Choi, J. Y. (2014). Audio bank: A high-level acoustic signal representation for audio event recognition. In *Proceedings of the 14th International Conference on Control, Automation and Systems (ICCAS)* (pp. 82–87). doi:10.1109/ICCAS.2014.6987963.

Schroder, J., Goetze, S., & Anemuller, J. (2015). Spectro-temporal Gabor filter-bank features for acoustic event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*, 2198–2208. doi:10.1109/TASLP.2015.2467964.

Schuller, B., Weninger, F., Wollmer, M., Sun, Y., & Rigoll, G. (2010). Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 4562–4565). doi:10.1109/ICASSP.2010.5495567.

Sebban, M., & Nock, R. (2002). A hybrid filter/wrapper approach of feature selection using information theory. *Pattern Recognition*, *35(4)*, 835–846. doi:10.1016/S0031-3203(01)00084-X.

Steeneken, H. J. M. (1991). Speech level and noise level measuring method. *Technical Report. Document SAM-TN0-042. Esprit-SAM*, (pp. 1–20).

Temko, A., & Nadeu, C. (2006). Classification of acoustic events using SVM-based clustering schemes. *Pattern Recognition*, *39*, 684–694. doi:10.1016/j.patcog.2005.11.005.

UPC-TALP (2012). UPC-TALP database of isolated meeting-room acoustic events. ELRA Catalog no. S0268.

Wilson, K., Raj, B., Smaragdis, P., & Divakaran, A. (2008). Speech denoising using nonnegative matrix factorization with priors. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 4029–4032). doi:10.1109/ICASSP.2008.4518538.

Zhang, Z., & Schuller, B. (2012). Semi-supervised learning helps in sound event classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 333–336). doi:10.1109/ICASSP.2012.6287884.

Zhuang, X., Zhou, X., Hasegawa-Johnson, M. A., & Huang, T. S. (2010). Real-world acoustic event detection. *Pattern Recognition Letters*, *31*, 1543–1551. doi:10.1016/j.patrec.2010.02.005.

Zhuang, X., Zhou, X., Huang, T. S., & Hasegawa-Johnson, M. (2008). Feature analysis and selection for acoustic event detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 17–20).

Zieger, C. (2008). An HMM based system for acoustic event detection. In *Lecture Notes in Computer Science (LNCS), Springer* (pp. 338–344). volume 4625. doi:10.1007/978-3-540-68585-2_32.