J Intell Inf Syst
DOI 10.1007/s10844-017-0481-4

Environmental sound recognition using short-time feature aggregation

Gerard Roma[1] · **Perfecto Herrera**[2] · **Waldo Nogueira**[3]

Received: 29 March 2017 / Revised: 26 July 2017 / Accepted: 28 July 2017

Abstract Recognition of environmental sound is usually based on two main architectures, depending on whether the model is trained with frame-level features or with aggregated descriptions of acoustic scenes or events. The former architecture is appropriate for applications where target categories are known in advance, while the later affords a less supervised approach. In this paper, we propose a framework for environmental sound recognition based on blind segmentation and feature aggregation. We describe a new set of descriptors, based on Recurrence Quantification Analysis (RQA), which can be extracted from the similarity matrix of a time series of audio descriptors. We analyze their usefulness for recognition of acoustic scenes and events in addition to standard feature aggregation. Our results show the potential of non-linear time series analysis techniques for dealing with environmental sounds.

Keywords Audio databases · Event detection · Environmental sound recognition · Audio features · Recurrence quantification analysis · Pattern recognition

✉ Gerard Roma
gerard.roma@gatech.edu

Perfecto Herrera
perfecto.herrera@upf.edu

Waldo Nogueira
nogueiravazquez.waldo@mh-hannover.de

[1] School of Literature, Media and Communication, Georgia Institute of Technology, Atlanta, GA, USA

[2] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

[3] Department of Otolaryngology, Medical University Hannover and Cluster of Excellence Hearing4all, Hannover, Germany

Published online: 19 August 2017

1 Introduction

Technologies for recording digital audio at a reasonable quality have become increasingly available and affordable. As a result, it is now very easy to collect information from our environment through audio recording. However, manual segmentation and labeling of audio are labor-intensive tasks. Thus, potential uses of that information are limited by our ability to automatically describe the contents of the recordings, which is still quite modest. In this paper, we address two problems that have received considerable attention in recent years: classification of acoustic scenes, and automatic labelling of acoustic events. In the first case, long recordings have to be assigned to one type of scene or environment, such as a train station or an office. In the second case, we zoom into the particular events that happen in scenes and attempt to detect and annotate their sources, such as human voices or alarms. An evaluation methodology and associated datasets were proposed in the context of the first IEEE AASP Detection and Classification of Acoustic Scenes and Events (*D-CASE*) Challenge (Giannoulis et al. 2013). In this paper we present an evolution of our work submitted to this challenge (Roma et al. 2013), and validate it with additional experiments and datasets. During the preparation of this work, a second *D-CASE* challenge has been conducted, and thus we will refer to the first one as *D-CASE2013*, and to the second one as *D-CASE2016*.

1.1 Related work

1.1.1 Acoustic scenes

The idea of classifying acoustic scenes into discrete categories has met growing interest with potential applications in robotics (Chu et al. 2006) and ubiquitous computing (Eronen et al. 2006). Most approaches can be classified into two groups: in the first case, some descriptors are computed frame-wise from the audio signal using short overlapping windows. The most common features are Mel Frequency Cepstral Coefficients (MFCCs). Some statistical model is then trained directly from the frame-level features, without any integration step. Scenes are commonly considered to be mainly stationary (Chachada and Kuo 2014; McDermott and Simoncelli 2011), so their identification does not depend on the long-term temporal evolution of audio features. Thus, Gaussian Mixture Models (GMMs) are a common choice (Aucouturier et al. 2007; Dargie 2009), although low-order Hidden Markov Models (which become a GMM in the single-state case) have also been used (Eronen et al. 2006). These methods are useful when the classes are known *a priori*, and can be applied in an online (e.g. real-time) setting.

The second approach consists in summarizing the entire recording into a single feature vector, often ignoring the order (i.e. a "bag of frames") and training the classification model document-wise with these vectors. Again, given the stationarity assumption it is common to use simple statistics of MFCC features. Support Vector Machines (SVM), K-nearest neighbor (KNN) and GMMs are popular choices for the classification model (Chu et al. 2006). Other than traditional MFCC features, promising results (although very different depending on the class) have been obtained with matching-pursuit features (Chu et al. 2009). However, their computational cost may make them unpractical in the context of large audio databases. Methods based on vector-quantization of frame-level descriptors have also been tried with acoustic scenes in the context of consumer videos (Lee and Ellis 2010), although the improvement over standard statistics is small. In this case, a global dictionary is first obtained by clustering a large amount of frames. Each file is then represented as a

normalized histogram of the dictionary items. One problem with this method is that the dictionary is often trained with the whole dataset, before partitioning it into training and test sets. This avoids re-training the dictionary for each fold when cross-validation is used for evaluation, but favors overfitting. Methods based on aggregation can be used for labelling using different vocabularies in larger scale applications, for instance when recordings are stored in a database. More recently, Non-negative matrix factorization (NMF) techniques have been introduced for unsupervised feature learning in scene classification (Bisot et al. 2016). Another related proposal is the use of latent acoustic topics, learnt from event signals (Imoto et al. 2013).

One general problem in comparing different methods is the lack of common datasets and benchmarks. The *D-CASE2013* challenge was proposed and conducted for benchmarking different methods for recognition of acoustic scenes and events. The baseline system proposed by the challenge organizers was based on MFCCs and a GMM classifier, while the best performing systems used SVMs. The highest accuracy for scene classification was achieved by the system described in this article, summarizing the evolution of MFCCs in intermediate windows of 400 ms. Another approach achieved similar performance using a wavelet variant of MFCCs and other features averaged over windows of 4 s, plus an extra classifier to decide for the class of the whole recording (Geiger et al. 2013). Such intermediate "texture windows" are commonly used in Music Information Retrieval (MIR) (Tzanetakis and Cook 2002). In the second iteration of the *D-CASE* challenge (*D-CASE2016*) some of the best approaches used deep learning architectures. These architectures are being used in many domains because of their ability to generalize using large training datasets (at high computational cost). However, research datasets in this field are still small.

1.1.2 Acoustic events

When we focus on acoustic events they usually happen in an environment that has a characteristic acoustic signature, usually as background "noise" (i.e., a bird chirp over a park environment background, a train whistle over a railway station background, etc.). We can therefore decompose those sounds into background noise and salient acoustic events. With respect to events, two main tasks can be identified: detection of events and their classification into discrete categories.

The task of segmentation, that is, finding the location of acoustic events in longer recordings, has been approached from several disciplines. In speech recognition, effective techniques have been developed for Voice Activity Detection (VAD) (Sohn et al. 1999), which have been widely applied in telecommunications. In MIR, a long tradition exists for the detection of instrument note onsets (Klapuri 1999; Böck and Widmer 2013). From the perspective of environmental sound, systems have been developed for applications such as surveillance (Clavel et al. 2005), or indexing of video content (Xu et al. 2003).

The task of acoustic event classification can be traced back to early experiments with audio retrieval, mainly in the context of indexing sound effects libraries. Early research focused on small datasets assigned to a handful of concepts. Very high classification accuracies were reported using HMMs (Zhang and Kuo 1998). For big collections of sound effects, though, it is more common to aggregate features at the file level and use generic machine learning models. For example, statistics of a large set of features have been used along with K-NN classifiers for large scale applications (Cano et al. 2005).

Like in the case of scenes, vector quantization has been used as a simple yet effective way to obtain sparse features (Chechik et al. 2008; Lee and Ellis 2010). Also, like in

the case of scenes, most recent work still oscillates between the HMM/GMM paradigm (Heittola et al. 2013) and the feature aggregation approach (Huang et al. 2013). The *D-CASE2013* challenge included tasks for identification of both overlapping and non-overlapping events. Most approaches were based on some sort of HMM, including the top performers. In this paper, we show how a combination of blind segmentation and SVM classification approaches the best results obtained in the challenge. Again, HMMs are better suited for joint segmentation and annotation of a few predefined concepts, but for large scales, including many files and labels, their complexity makes them less attractive. Like in the case of scenes, deep learning architectures are becoming increasingly popular for event detection (Zhang et al. 2015), but their computational cost makes them also less attractive for large vocabularies. Aggregation-based systems afford a more generic and cost-effective approach, as the number of required parameters is generally smaller.

1.1.3 Feature aggregation

In this paper, we analyze the problem of feature aggregation for environmental sound. Aggregation of features over time is usually a required step when using many machine learning algorithms to automatically label audio segments or files. Most approaches are driven by results on available research data, with limited theoretical insight about the temporal evolution of the signal. An example is automatic generation of features (Pachet and Roy 2007). More recently data-driven unsupervised feature learning has become popular (Lee et al. 2009). Feature aggregation based on theoretical insight is less common. A common approach in MIR is to consider musical concepts such as rhythm (Tzanetakis and Cook 2002). In the case of environmental sound, feature aggregation can still be considered an open issue. As an example, the "bag of frames" model associated with common statistics was initially considered "good enough" (Aucouturier et al. 2007), but the merits were more recently attributed to the research datasets used (Lagrange et al. 2015). While speech and conventional musical audio is produced by well known specialized sound production systems (typically the vocal tract and musical instruments), environmental audio can contain a wide range of sounds. For instance, the datasets used in the scene classification experiments (Section 3.2) contain recordings from urban environments, such as public transportation, quiet and busy streets, or restaurants. Many sounds in these environments are produced by natural and social phenomena typically studied as complex dynamic systems, often exhibiting chaotic behaviour, such as weather conditions (wind, rain ...) or car traffic. RQA features were devised to describe this kind of systems and have been applied in many different domains. With respect to audio, RQA features have been applied to pathological voice using raw waveforms (Washington et al. 2012) and only introduced for environmental sound in the prototype version of this system, evaluated for acoustic scene classification (Roma et al. 2013). RQA features derived from frame-level chroma features have been tested in the cross-recurrence setting, where two different series are compared, for cover song detection (Serrà et al. 2009). A more similar work has tested recurrence time histograms (in this case extracted directly from the audio signal) for genre classification (Serrà et al. 2011).

We propose an integrated framework for environmental sound recognition based on blind segmentation and feature aggregation, and evaluate the use of RQA features for classification of acoustic scenes and events. In the next section, the main components of the framework are described. In Section 3, we present several experiments to assess different algorithms for each component.

2 Proposed approach

2.1 Environmental sound recognition framework

Considering aggregation of both sound scenes and events affords a general perspective on the issue of environmental sound recognition. This way, we can analyze our environment as a composition where events are perceived in the context of a sound scene. By combining event detection and feature aggregation, a general framework for automatic recognition of environmental sound can be devised and implemented. Figure 1 shows a block diagram of the proposed framework. The system is based on blind segmentation of monophonic events, implying a distinction between salient events and background. This allows more flexible and scalable architectures, since the classifier models do not need to be evaluated for all of the audio in a scene. The combination of scene classification and event detection and recognition obviously depends on the development of suitable categories. While the issue of concept ontologies is beyond the scope of this paper, an example is provided by the "office" category in the *D-CASE2013* challenge. In an application scenario, a scene classification model could be used to recognize "office" scenes (Section 3.2) and then a segmentation and classification step would be used to detect events happening in an office (Section 3.3).

The process starts by performing a time-frequency analysis via the Short Time Fourier Transform (STFT) (note that other time-frequency representations are possible without changing the general scheme). For scene classification, spectral frames are parametrized using MFCCs (again other descriptors could be used). This results in a multivariate time series that is aggregated at two levels: first we compute general statistics of the whole available recording of the scene. We analyze also short-term texture windows, from which we extract both conventional statistics and non-linear features. Local features are averaged and concatenated with global statistics. The joint feature vector is then classified by an SVM model, which is trained with previously labelled scenes.

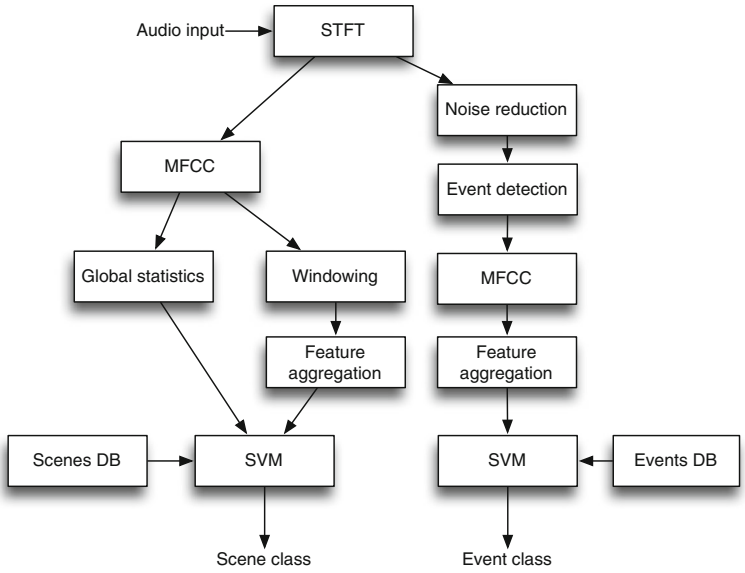

Fig. 1 Block diagram of the proposed system

For event recognition, several algorithms can be used for event detection from the spectral frames. Before segmentation, a noise reduction stage is introduced, since events are embedded in the scene background. Frames within each segment are then aggregated directly (i.e. no texture window is used) and the segments are classified in the same way as scenes.

2.2 Noise reduction

The assumption of salient events over background noise is common for speech processing. Single channel noise reduction (SCNR) algorithms are commonly used for speech enhancement (Martin 1994), for example they are used in automatic speech recognition systems (Yu et al. 2008) and hearing aid devices (Brons et al. 2014) to improve speech recognition performance. The same principles can be applied for environmental sound.

The SCNR algorithm is applied to the event detection subsystem from the initial STFT frames $Y(n)$ of the incoming sound y(l), where n and l denote the frame index and the discrete time index respectively. It is assumed that the noise is additive, i.e. $y(l) = x(l) + v(l)$, where $x(l)$ denotes the acoustic events and $v(l)$ the additive noise. The algorithm is applied to a set of spectral bands combined from the linearly-spaced FFT bands. The bands are spaced logarithmically along frequency with constant Q. Each band z contains L_z bins and so the power in each band is computed as follows:

$$A_y^2(n,z) = \sum_{k=k_{start_z}}^{k_{start_z}+L_z} Re\{Y^2(n,k)\} + Im\{Y^2(n,k)\}, \tag{1}$$

where k_{start_z} denotes the FFT bin number where band z starts. Moreover, it is assumed that the acoustic events and the noise are uncorrelated, i.e.:

$$E\{A_y^2(n,z)\} = E\{A_x^2(n,z)\} + E\{A_v^2(n,z)\}, \tag{2}$$

where $A_x^2(n,z)$ and $A_v^2(n,z)$ denote the power of the acoustic event and the background noise signals respectively.

Given the signal model, a subtraction rule for the noise reduction filter $H(n,z)$ is derived by minimizing the mean-squared error:

$$H(n,z) = \underset{H(n,z)}{argmin}\ E\{\|A_x(n,z) - H(n,z)A_y(n,z)\|^2\}. \tag{3}$$

The result is the well-known Wiener filter:

$$H(n,z) = \frac{E\{\|A_x^2(n,z)\|\}}{E\{\|A_x^2(n,z)\|\} + E\{\|A_v^2(n,z)\|\}}. \tag{4}$$

In order to implement the filter it is necessary to estimate $\tilde{A}_x^2(n,z)$ and $\tilde{A}_v^2(n,z)$ from the noisy observation $A_y^2(n,z)$.

The SNR, i.e. the difference between the event and noise power in dB, is estimated assuming that events produce fast increases and decreases in power with respect to the background noise. For this reason, the event $\tilde{A}_x^2(n,z)$ and the background noise $\tilde{A}_v^2(n,z)$ power are estimated using first-order recursive averaging with a relatively fast and slow time constant respectively:

$$\tilde{A}_x^2(n,z) = \alpha_s \tilde{A}_x^2(n,z) + (1-\alpha_s)A_y^2(n,z), \tag{5}$$

$$\tilde{A}_v^2(n,z) = \alpha_v \tilde{A}_v^2(n,z) + (1-\alpha_v)A_y^2(n,z), \tag{6}$$

where $\alpha_s = e^{\frac{1}{\tau_s}}$ and $\alpha_v = e^{\frac{1}{\tau_v}}$ are the smoothing constants for the events and the background noise respectively. When $\tilde{A}_x^2(n,z)$ exceeds $\tilde{A}_v^2(n,z)$ by 6 dB it is assumed that an event occurred and (5) is updated, otherwise (6) is updated. For each audio frame, $\tilde{A}_x^2(n,z)$ and $\tilde{A}_v^2(n,z)$ are substituted in (4) to obtain the noise reduction filter.

2.3 Event detection

Event segmentation or event activity detection (EAD) can be used as an independent step for event recognition. In this section, we analyze approaches for monophonic event detection, i.e. only one event is recognized at a given time. While research on polyphonic event detection in environmental sound is ongoing (Parascandolo et al. 2016), there are still valid use cases for monophonic detection in environmental sound, such as application to quiet environments (home, office) or detection of specific events. As an example, detecting crowd excitement in sports audio is a polyphonic event that could be detected by a monophonic method as a general energy transient. In this paper, we analyze the usefulness of common monophonic event detection methods, developed in speech and music, for environmental sound recognition. Voice activity detection (VAD) attempts to detect speech in the context of relatively stable background noise. In the simplest case, detection is based on abrupt energy changes, while more sophisticated approaches take into account specific characteristics of speech. Here, we deal with different kinds of sounds in the context of a sound scene. Thus, generic VAD approaches based on short-time energy could be useful. Similarly, onset detection functions commonly used in MIR can be used to detect salient events in environmental audio in a generic way. Here, we consider the ability of these functions to identify the whole event instead of only the onset as is typical in MIR. Perhaps the most basic one would be raw spectral energy:

$$E(n) = \sum_{k=0}^{N-1} |X_n(k)|^2, \tag{7}$$

where $X_n(k)$ denotes the spectral band k at time frame n. High frequency content (HFC), denotes a frequency dependent weight for spectral energy, which usually works for percussive (i.e. non-pitched) onsets in music:

$$HFC(n) = \sum_{k=0}^{K-1} k|X(k)|^2. \tag{8}$$

We also consider a version (denoted HFC_mel in the experiments) computed over the mel bands in the MFCC framework, in the spirit of the variant proposed by Scheirer (1998). Another common function is the "spectral flux", computed as the magnitude of the complex difference between successive frames:

$$FLUX(n) = \sum_{k=0}^{N-1} |X_n(k) - X_{n-1}(k)|. \tag{9}$$

The ITU (ITU-T 2010) algorithm indicates if the input frame is speech or music (speech/music discrimination). For an inactive frame, it indicates whether the frame is a silence frame or an audible noise frame (silence detection). This algorithm has been used to detect transitions from one state to another, therefore allowing its use also for the segmentation of events.

2.4 Recurrence quantification analysis features

After blind segmentation (using texture windows in the case of scenes, or using event detection functions in the case of events), we compute feature agregation, notably RQA features. RQA (Zbilut and Webber 2006) is a set of techniques developed during the last decade in the study of chaos and complex systems. The basic idea is to quantify patterns that emerge in recurrence plots, which are obtained by thresholding the self-similarity matrix of a time series. RQA has been applied in a wide variety of disciplines. The original technique starts from one-dimensional time series which are assumed to result from a process involving several latent variables. This multidimensionality is recovered by delaying the time series and embedding it in a phase space. The distance matrix of the series is then computed and thresholded to a certain radius r. The radius represents the maximum distance of two observations of the series that will still be considered as belonging to the same state of the system.

In the case of audio analysis, it is common to work with multivariate spectral time series. We adapt the technique by computing and thresholding the similarity matrix obtained from the MFCC representation using a cosine distance. This distance metric is usually preferred to Euclidean distance for high-dimensional spaces and provided better results in early experiments.

If we denote the series of cepstral coefficients vectors as $C = C_1, C_2, C_3...C_N$, then the recurrence plot R is defined as

$$R_{i,j} = \begin{cases} 1 \text{ if}(1 - \frac{C_i \cdot C_j}{||C_i|| ||C_j||}) < r \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

where r is the radius, and i, j are time indices.

Figure 2 shows two of such plots, one (left side) corresponding to a sound scene recorded in a subway train (with high recurrence levels) and the other (right side) to a scene recorded in a restaurant (low recurrence but some straight lines). The most relevant information comes from diagonal lines, which represent periodicities in the signal (i.e., repeated or,

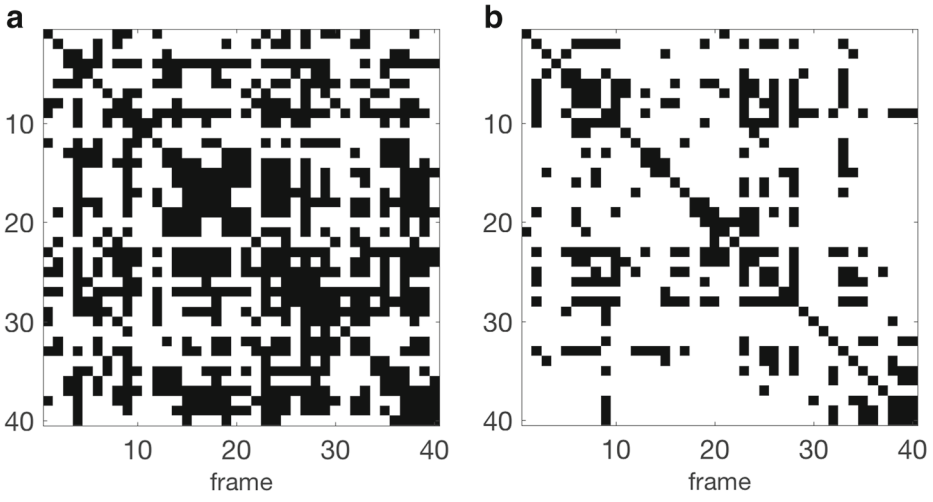

Fig. 2 Recurrence plots of *tube* (higher recurrence) and *restaurant* (lower recurrence) acoustic scenes, using windows of 40 frames

depending on the radius, quasi-repeated sequences of frames), and from vertical lines (or horizontal, since the plot is symmetric), which represent stationarities, i.e., the system remains in the same state. From these representations, several metrics have been developed to quantify the size and length of lines of contiguous points in the matrix. Most features were developed by Webber and Zbilut (1994). We extract the most commonly used ones and add some more variables in order to obtain more features for the classifier. For completeness, equations are included in the Appendix.

In order to analyze long series, a windowed version is often used, which consists in computing the recurrence plots from overlapping windows of fixed size, as opposed to using the whole time series. This is computationally much more efficient, and we found it to give similar results in early experiments with acoustic scenes. Our interpretation of this is that, in the case of scenes, relevant recurrences happen at a short time scale. The underlying assumption is that the recurrence plots allow us to identify patterns in background noise texture. Regarding the radius parameter, a set of guidelines for choosing a suitable value are found in Webber and Zbilut (2005). Mainly, recurrence rate (see the Appendix) follows a sigmoid curve as the radius is increased, and the value is chosen within the scaling region of the sigmoid. We found the same behaviour for the case of audio spectral features, and chose a value of $r = 0.6$ that works for all datasets in this paper, including scenes and events. This ensures that our choice does not overfit a particular dataset. Similarly, for all scenes datasets, a window size of 1.28 seconds was used. Features were standardized for each window, which allowed the use of longer windows than in Roma et al. (2013) (note that this also affects the useful range of r with respect to non-standardized windows). Figure 3 illustrates the main steps in the computation of the recurrence plot from the spectrogram.

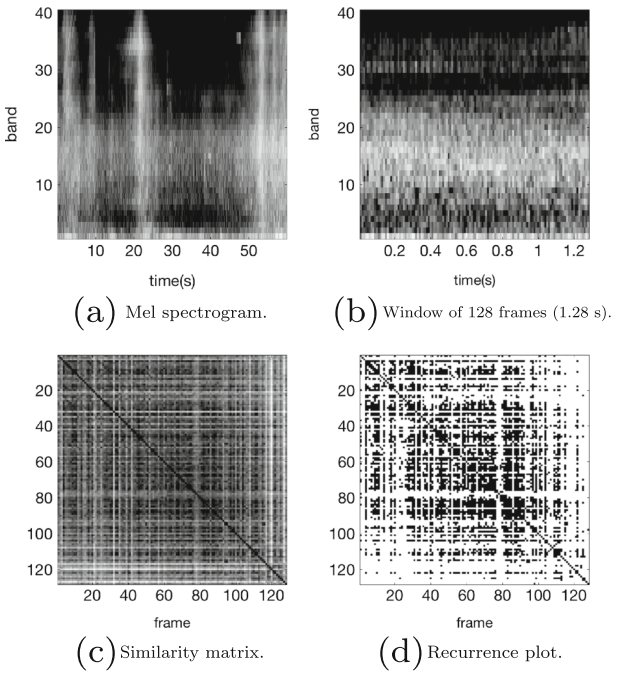

Fig. 3 Steps in the construction of the recurrence plot

3 Experiments

3.1 Overview

We have proposed a general framework for computational analysis of environmental sound. Our approach is based on feature aggregation and classification of acoustic scenes, and on blind segmentation, aggregation and classification of acoustic events. In this section we compare different strategies and feature sets for the main components of the described architecture. The first experiment compares different aggregation features for acoustic scene classification. Analogously, the second experiment compares aggregation features for event classification. In the third experiment, several methods for segmenting scenes into events are compared. In a fourth experiment, the results of the previous two are analyzed in the context of a joint segmentation and classification task, including the effect of noise removal.

The general methodology for the different experiments is as follows: first, STFT frames were extracted from each recording using a sliding window of 25 ms with hops of 10 ms. MFCCs (13 coefficients from 40 bands) were computed using the rastamat (Ellis 2005) library. A larger number of coefficients provided no improvement in preliminary experiments, while a smaller number had a negative impact, so the default was used. The resulting frame-level features were aggregated using common statistics. From the same features, RQA descriptors were obtained using recurrence plots computed via cosine distance (see the Appendix). The total number of features is detailed in each experiment. Annotation of unseen audio was done through an SVM classifier. We used the libSVM (Chang and Lin 2001) implementation with the default RBF kernel, the main hyperparameters being the classification cost C and the kernel parameter γ. Evaluation was done using 5-fold cross-validation following the *D-CASE2013* methodology. A nested cross-validation step (Cawley and Talbot 2010) was used to select hyperparameters using grid search for each fold of the outer cross-validation loop. The process was repeated for 100 times for different feature aggregation sets, in order to account for the random partition in folds (Alpaydin 2014). The resulting distributions were compared using a pair-wise Wilcoxon signed-rank test. Specific settings for each experiment are described in the following sections.

3.2 Features for scene classification

Many acoustic scenes show low variability along a considerable time span (e.g. park atmosphere, background in a tube station when the train is not there, beach seashore), and in this sense, they can be characterized from global statistics. At the same time, the texture of the background noise is likely to be an important cue for human recognition of different environments, such as urban soundscapes. In this experiment, RQA features were compared to standard aggregation (bag of frames). RQA features do not require assumptions about linearity or stationarity of the time series. Hence, they seem a good complement or alternative for the traditional averaging of spectral audio features in the case of environmental audio, as in many environments there happen to be recurrent events.

3.2.1 Method

In order to test the generalization of the method proposed in Roma et al. (2013) beyond the challenge dataset, we implemented classification of sound scenes using three different datasets: the *D-CASE2013* dataset (Giannoulis et al. 2013) contains 100 recordings of 30s, corresponding to 10 different scenes (bus, busy street, office, open air market, park, quiet

street, restaurant, supermarket, tube and tube station). The *in-house* dataset (Roma et al. 2013) contains recordings of the same categories (150 recordings, 15 per class) but collected from several collections, mainly sound effects CDs or web pages. Finally, the *Rouen* dataset (Rakotomamonjy and Gasso 2015) contains 3026 recordings of 19 different classes (plane, bus, busy street, cafe, car, hall, *hall gare*, kid game, market, metro-paris, metro-rouen, pool hall, quiet street, restaurant, *rue pietonne*, shop, train-ter, train-tgv and tube-station).

For the case of scenes, all files were analyzed as described in Section 3.1 with a frequency threshold of 5000 Hz[1] using a texture window of 1.28s. In addition to general statistics for each recording, we also computed the local variance of features in the texture window, and averaged all of these local features over the recording. In this context, we compared classification accuracy using different feature aggregation sets: global mean and variance (*mv*), adding local variance (*mvlv*), adding RQA (*mvrqa*), and adding both (*mvlvrqa*). The overall set of features (*all*), computed over 13 cepstral coefficients, amounts to 50 dimensions for describing sound scenes.

3.2.2 Results and discussion

Figure 4 shows the results for each set of features for the three scenes datasets. All results are well above the 10% random baseline (5% for the *rouen* dataset), but still below the median human recognition rate of 75% reported for the *D-CASE2013* dataset (Barchiesi et al. 2015) (again except for *rouen*). It can be seen that local statistics improve accuracy in all cases. Also, *lv* and *rqa* increase the overall accuracy when added together, which indicates that they provide complementary information. Two of the datasets in this experiment, *D-CASE2013* and *rouen*, were recorded in uniform conditions, i.e. using the same equipment, while the *in-house* dataset contains sounds from different collections. Both the local variance and *rqa* features, and also their combination, produce a greater improvement when combined with global statistics. In general, our approach based on feature aggregation performs similarly to models trained frame-by-frame for specific concepts (Stowell et al. 2015), but it can also be used for analyzing databases with uncontrolled recording conditions. The best results are obtained using the *rouen*, by far the largest of the three, which shows a clear improvement for each descriptor set. Our results for this dataset are better than the MFCC+RQA combination reported in Rakotomamonjy and Gasso (2015) (86% accuracy) but probably below more recent, though not directly comparable results (94.5% F1-score in Bisot et al. 2016).

3.3 Features for event classification

Like in the case of sound scenes, we compared RQA features to common methods for aggregation of the time series of MFCCs.

3.3.1 Method

We analyzed the classification performance of an SVM classifier using different sets of features and two different datasets. The first dataset was provided as part of the *D-CASE2013*

[1]Note that our original submission to the *D-CASE* challenge obtained very good results by narrowing the frequency range to below 1 Khz. However in our experiments this only worked for the challenge dataset, so for the sake of generality here we chose a wider range.

Fig. 4 Classification accuracy for scene classification. All pair-wise differences are statistically significant ($p < 0.05$) except when noted (ns = not significant). *Boxes* extend from lower to upper quartile, *whiskers* indicate highest and lowest points within 1.5 IQR, points indicate outliers

challenge as a training set for the event detection task. It contains 320 samples classified into 16 classes. The second dataset was collected in a previous work when investigating audio indexing using the sound events taxonomy by Gaver (1993). It contains 1608 samples from various collections, divided into 11 classes.

Apart from traditional statistics, MFCC and RQA features were computed as described in Section 3.1. The following descriptor sets were compared: mean and variance (*mv*) (26 dimensions), adding mean and variance of the derivative (*mvdmdv*, 52), adding RQA (*mvrqa* 37) and adding both (*all*, 63).

3.3.2 Results and discussion

Figure 5 shows the classification accuracy for each set of features with the two events datasets. Both results are also well above the implicit random baseline (6% and 9% respectively). In this case the samples in the *d_case_events* dataset were all recorded in similar conditions, while the *gaver_events* contains sounds from different collections. RQA features tend to give larger improvements than feature derivatives. This is especially noticeable for the *gaver_events*, which contains a much larger amount of sounds. Consistently with the previous experiment using scenes, our results confirm that using short-time feature aggregation improves classification accuracy.

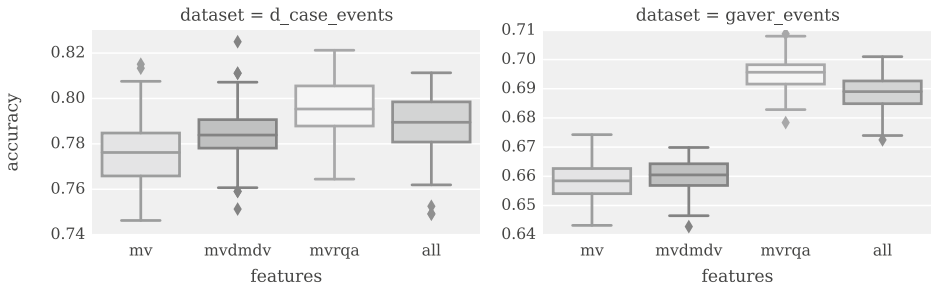

Fig. 5 Classification accuracy for event classification. All pair-wise differences are statistically significant ($p < 0.05$). *Boxes* extend from lower to upper quartile, *whiskers* indicate highest and lowest points within 1.5 IQR, points indicate outliers

3.4 Event segmentation

In this experiment, we compare two VAD implementations denoted VAD (Sohn et al. 1999) and ITU (ITU-T 2010) in addition to onset detection functions discussed in Section 2.3: HFC, HFC_mel, FLUX and E (raw energy).

3.4.1 Method

All the onset detection functions (except VAD and ITU) were post-processed in the following way: first, the function was z-scored, then smoothed with a 5-point moving average filter, and then a long term (400-point) moving median filter was subtracted from the resulting signal (Bello et al. 2005). The VAD and ITU algorithms give a continuous probability estimate and were not post-processed.

The algorithms were evaluated using the *event office live* dataset from the *D-CASE2013* challenge, composed of three scripted recordings with labelled non-overlapping events. For each recording, we computed a binary signal denoting the spectral frames attributed to acoustic events vs background, and computed the Jaccard distance to the ground truth binary signal:

$$D_{jac} = 1 - \frac{p}{p+q+r} \tag{11}$$

where p is the number of frames where both the ground truth annotation and the estimate predict an event, and q, r are the number of non-zero frames only in either the ground truth or the estimate respectively.

3.4.2 Results and discussion

Figure 6 shows boxplots of the Jaccard distance when using each detection function for the three scripts of the dataset.

It can be observed that the VAD and the HFC_mel functions perform best. However, it should be noted that the dataset, while not synthetic, was curated to provide an easy task, also characterized by low noise conditions (indoor scenes). In this situation, the described assumptions of the VAD algorithm about background noise do hold. In more realistic situations, this algorithm may result in many false positives.

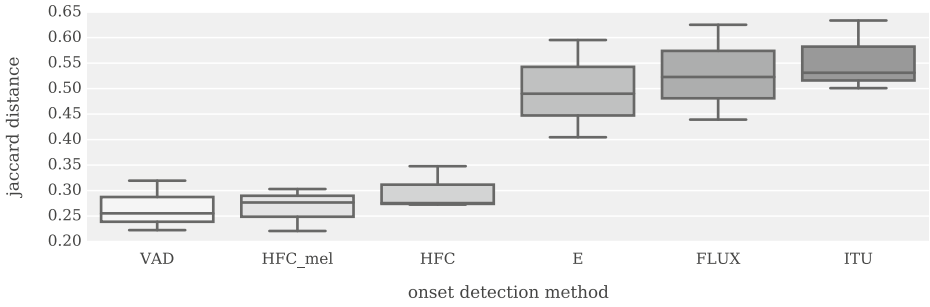

Fig. 6 Results for different onset detector methods. The results are given based on the Jaccard distance, the lower this value, the better the performance. *Boxes* extend from lower to upper quartile, *whiskers* indicate highest and lowest points within 1.5 IQR

3.5 Joint detection and classification

In previous experiments, we have compared methods for blind segmentation (i.e., where no model is used for the nature of the detected event) and for classification of sound events. The idea was to provide a flexible framework that can be used in large-scale applications, as opposed to detecting events by training models for specific classes of sounds. We now compare the best features and event detection functions in a joint event recognition task (i.e., detecting that there is an event and deciding which category it belongs to). We also evaluate the use of the noise reduction algorithm described in Section 2.2.

3.5.1 Method

Again the methodology and data for this experiment followed the event detection task in the *D-CASE2013* challenge. Using the same scripted recordings than in the segmentation experiment, the EAD functions were computed to identify segments corresponding to events, which were classified using the same SVM model for the *d_case_events* dataset in experiment Section 3.3. For the noise reduction step, the parameters τ_s and τ_v (see Section 2.2) were chosen experimentally to optimize the recognition of the events with values of 10 and 2000 ms respectively. The training set of event sounds was recorded in better conditions and did not require noise removal, while the test scene contained consistent background noise. Thus, SCNR was only applied to the test dataset. In the test stage, we considered whether each 10 ms frame had been correctly assigned to its labelled class (note that many different evaluation measures were used in the challenge (Giannoulis et al. 2013), we report only frame-level F-measure for simplicity).

3.5.2 Results and discussion

Figure 7 shows the classification results for the feature sets of Section 3.3 and the best segmentation functions of Section 3.4 (including ideal segmentation where the ground truth annotations are used only for the segment boundaries) with and without noise reduction. Unlike the plain classification task, derivative features generally do not improve classification, while RQA features provide small improvements. Since the differences with the

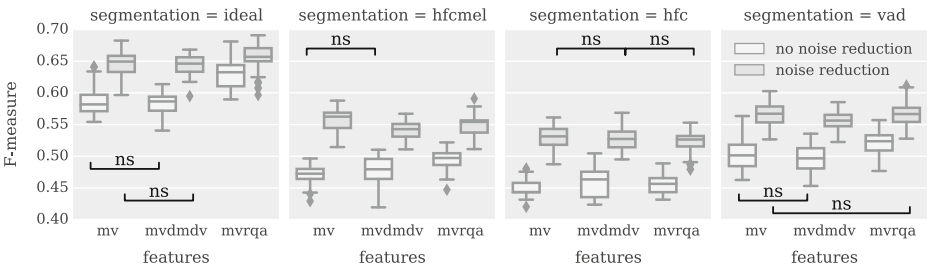

Fig. 7 Results for event recognition using different feature sets and segmentation functions, with or without noise reduction. All pair-wise differences across feature sets are statistically significant ($p < 0.05$) except when noted (ns = not significant). All pair-wise differences across segmentation functions and across the noise reduction condition are statistically significant ($p < 0.05$). *Boxes* extend from lower to upper quartile, *whiskers* indicate highest and lowest points within 1.5 IQR, points indicate outliers

classification task also hold in the ideal case, they are not due to segmentation errors, but most probably to the presence of background noise in the recorded scripts. The inclusion of the noise reduction step improved recognition in all cases. In that case the effect of adding RQA features is no longer noticeable, except for ideal segmentation. This seems to imply that these features are less disturbed by background noise, but are influenced by segmentation errors. Using noise reduction and VAD for event detection, our system achieved a value of 56% for frame-level F-measure, which is second to the best (61.52%) but above the rest of systems submitted for the event recognition task of the *D-CASE2013* challenge (Stowell et al. 2015). It should be noted that in this experiment uses a relatively small test set (in the order of 30 events per script, results being averaged across three scripts) where the classes are not balanced. Thus, the experiment provides limited insight. In contrast, the *gaver_events* dataset used in Section 3.3 contains more than 1500 instances.

4 Conclusions

Recognition of environmental sound is a promising but also challenging research problem with applications in diverse areas, from ubiquitous computing to robotics. Many research efforts have, until now, concentrated on specific classes of sounds. This resulted in tailored models that are not necessarily suited for large-scale recognition. In this work, we have proposed a general framework for identifying arbitrary sounds both at the sound scene and the sound event level, including detection of events along time. By separating segmentation and classification, this architecture provides large-scale applications with more flexible ways for automatically annotating scenes and events. We have analyzed the most important components in this framework, with a focus on improving aggregation of frame-level spectral features.

With respect to segmentation, we have shown that onset detection functions adapted from MIR research can compete with traditional VAD on the task of blind event detection in controlled noise environments. Further work is required for generalizing this comparison to the case of events in noisy environments.

With respect to feature aggregation, we have introduced a set of features (RQA) from nonlinear time series analysis, that provide distinct improvements with respect to traditional statistics. We have shown that this features can be added to general statistics for improved classification accuracy. Also, they are robust to varying recording conditions and background noise, which makes them attractive for dealing with large-scale and unstructured databases.

With respect to background noise, we have seen that introducing a state-of-the art SCNR algorithm from speech processing can be used to consistently improve event recognition in the context of our framework.

In this framework, numerous opportunities are open for improving the different components, such as replacing SVMs with more novel classifiers, or using different time-frequency audio representations.

Acknowledgements The first author was with the Music Technology Group at Universitat Pompeu Fabra for the main part of this work. The third author was with the Music Technology Group at Universitat Pompeu Fabra for part of this work. This work has been suported by the DFG cluster of excellence EXC 1077/1 “Hearing4all”.

Appendix: RQA features

This section details the equations used for computing this features, mostly derived from Webber and Zbilut (1994). Here, R refers to recurrence plot as described in Section 2.4.

- Recurrence rate (REC) is just the percentage of points in the recurrence plot.

$$REC = (1/N^2) \sum_{i,j=1}^{N} R_{i,j} \tag{12}$$

- Determinism (DET) is measured as the percentage of points that are in diagonal lines.

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{i,j=1}^{N} R_{i,j}} \tag{13}$$

  where $P(l)$ is the histogram of diagonal line lengths l.
- Laminarity (LAM) is the percentage of points that form vertical lines.

$$LAM = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)} \tag{14}$$

  where $P(v)$ is the histogram of vertical line lengths v.
- The ratio between DET and REC is often used. We also use the ratio between LAM and REC, so we define them as

$$DRATIO = N^2 \frac{\sum_{l=l_{min}}^{N} lP(l)}{(\sum_{l=1}^{N} lP(l))^2} \tag{15}$$

$$VRATIO = N^2 \frac{\sum_{v=v_{min}}^{N} vP(v)}{(\sum_{v=1}^{N} vP(v))^2} \tag{16}$$

- LEN and Trapping Time TT are the average diagonal and vertical line lengths.

$$LEN = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=l_{min}}^{N} P(l)} \tag{17}$$

$$TT = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=v_{min}}^{N} P(v)} \tag{18}$$

- Another common feature is the length of the longest diagonal and vertical lines. The inverse of the maximum diagonal (called Divergence) is also used. We use the inverse of both vertical and diagonal maximum lengths.

$$DDIV = \frac{1}{max(l)} \tag{19}$$

$$VDIV = \frac{1}{max(v)} \tag{20}$$

- Finally, the Shannon entropy of the diagonal line lengths is commonly used. We also compute the entropy for vertical line lengths.

$$DENT = -\sum_{l=l_{min}}^{N} P(l) ln(P(l)) \tag{21}$$

$$VENT = -\sum_{v=v_{min}}^{N} P(v)ln(P(v)) \tag{22}$$

References

Alpaydin, E. (2014). *Introduction to machine learning*. MIT Press.

Aucouturier, J.J., Defreville, B., & Pachet, F. (2007). The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, *122*(2), 881.

Barchiesi, D., Giannoulis, D., Stowell, D., & Plumbley, M.D. (2015). Acoustic scene classification: classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, *32*(3), 16–34.

Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M.B. (2005). A tutorial on onset detection in music signals. *IEEE Audio, Speech Language Processing*, *13*(5), 1035–1047.

Bisot, V., Serizel, R., & Essid, S. (2016). Acoustic scene classification with matrix factorization for unsupervised feature learning. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6445–6449). IEEE.

Böck, S., & Widmer, G. (2013). Maximum filter vibrato suppression for onset detection. In *Proceedings of the 16th international conference on digital audio effects (DAFx-13)*. Maynooth.

Brons, I., Houben, R., & Dreschler, W.A. (2014). Effects of noise reduction on speech intelligibility, perceived listening effort, and personal preference in hearing-impaired listeners. *Trends in Hearing* 18. https://doi.org/10.1177/2331216514553924.

Cano, P., Koppenberger, M., & Wack, N. (2005). An industrial-strength content-based music recommendation system. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (p. 673). Salvador.

Cawley, G.C., & Talbot, N.L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Chachada, S., & Kuo, C.C.J. (2014). Environmental sound recognition: a survey. *APSIPA Transactions on Signal and Information Processing*, *3*, e14.

Chang, C.C., & Lin, C.J. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chechik, G., Ie, E., Rehn, M., Bengio, S., & Lyon, D. (2008). Large-scale content-based audio retrieval from text queries. In *Proceedings of the 1st ACM international conference on multimedia information retrieval (MIR '08)* (p. 105). Beijing.

Chu, S., Narayanan, S., Kuo, C.C.J., & Mataric, M.J. (2006). Where am I? Scene recognition for mobile robots using audio features. In *2006 IEEE International conference on multimedia and expo* (pp. 885–888).

Chu, S., Narayanan, S., & Kuo, C.C.J. (2009). Environmental sound recognition with time-frequency audio features. *IEEE Audio, Speech Language Processing*, *17*(6), 1142–1158.

Clavel, C., Ehrette, T., & Richard, G. (2005). Events detection for an audio-based surveillance system. In: *IEEE International conference on multimedia and expo (ICME 2005)* (pp. 1306–1309).

Dargie, W. (2009). Adaptive audio-based context recognition. *IEEE Transactions on Systems, Man and Cybernetics Part A: Systems and Humans*, *39*(4), 715–725.

Ellis, D.P.W. (2005). PLP and RASTA (and MFCC, and inversion) in Matlab. http://www.ee.columbia.edu/ln/rosa/matlab/rastamat/. Online web resource.

Eronen, A., Peltonen, V., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G., & Huopaniemi, J. (2006). Audio-based context recognition. *IEEE Audio, Speech Language Processing*, *14*(1), 321–329.

Gaver, W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, *5*(1), 1–29.

Geiger, J.T., Schuller, B., & Rigoll, G. (2013). Recognising acoustic scenes with large-scale audio feature extraction and SVM. Tech. rep. IEEE AASP challenge: detection and classification of acoustic scenes and events.

Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., & Plumbley, M.D. (2013). Detection and classification of acoustic scenes and events: an IEEE AASP challenge. In *2013 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)* (pp. 1–4). IEEE.

Heittola, T., Mesaros, A., Eronen, A., & Virtanen, T. (2013). Context-dependent sound event detection. *EURASIP Journal on Audio Speech, and Music Processing*, *1*, 1.

Huang, Z., Cheng, Y.C., Li, K., Hautamäki, V., & Lee, C.H. (2013). A blind segmentation approach to acoustic event detection based on i-vector. In *Proceedings of interspeech* (pp. 2282–2286).

Imoto, K., Ohishi, Y., Uematsu, H., & Ohmuro, H. (2013). Acoustic scene analysis based on latent acoustic topic and event allocation. In *2013 IEEE international workshop on machine learning for signal processing (MLSP)* (pp. 1–6). IEEE.

ITU-T (2010). A generic sound activity detector recommendation G.720.1. https://www.itu.int/rec/T-REC-G.720.1/en.

Klapuri, A. (1999). Sound onset detection by applying psychoacoustic knowledge. In *Proceedings of 1999 IEEE international conference on acoustics, speech, and signal processing, 1999* (Vol. 6, pp. 3089–3092). IEEE.

Lagrange, M., Lafay, G., Defreville, B., & Aucouturier, J.J. (2015). The bag-of-frames approach: a not so sufficient model for urban soundscapes. *The Journal of the Acoustical Society of America*, *138*(5), EL487–EL492.

Lee, H., Pham, P., Largman, Y., & Ng, A.Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems* (pp. 1096–1104).

Lee, K., & Ellis, D.P.W. (2010). Audio-based semantic concept classification for consumer video. *IEEE Audio, Speech and Language Processing*, *18*(6), 1406–1416.

Martin, R. (1994). Spectral subtraction based on minimum statistics. *Proceedings of EUSIPCO*, *94*(1), 1182–1185.

McDermott, J.H., & Simoncelli, E.P. (2011). Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, *71*(5), 926–940.

Pachet, F., & Roy, P. (2007). Exploring billions of audio features. In *2007 international workshop on content-based multimedia indexing* (pp. 227–235). IEEE.

Parascandolo, G., Huttunen, H., & Virtanen, T. (2016). Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6440–6444). IEEE.

Rakotomamonjy, A., & Gasso, G. (2015). Histogram of gradients of time-frequency representations for audio scene classification. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *23*(1), 142–153.

Roma, G., Nogueira, W., & Herrera, P. (2013). Recurrence quantification analysis features for environmental sound recognition. In *2013 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)* (pp. 1–4). IEEE.

Scheirer, E.D. (1998). Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, *103*(1), 588–601.

Serrà, J., Serra, X., & Andrzejak, R.G. (2009). Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11.

Serrà, J., De los Santos, C., & Andrzejak, R.G. (2011). Nonlinear audio recurrence analysis with application to genre classification. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 169–172). IEEE.

Sohn, J., Kim, N.S., & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, *6*(1), 1–3.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. (2015). Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, *17*(19).

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Audio, Speech and Language Processing*, *10*(5), 293–302.

Washington, C.d.A., Assis, F.M., Neto, B.G.A., Costa, S.C., & Vieira, V.J.D. (2012). Pathological voice assessment by recurrence quantification analysis. In *2012 ISSNIP biosignals and biorobotics conference: biosignals and robotics for better and safer living (BRC)* (pp. 1–6). IEEE.

Webber, C.L., & Zbilut, J.P. (1994). Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology*, *76*(2), 965–973.

Webber, J.r., C. L., & Zbilut, J.P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences*, 26–94.

Xu, M., Maddage, N., Xu, C., Kankanhalli, M., & Tian, Q. (2003). Creating audio keywords for event detection in soccer video. In *Proceedings of the 2003 IEEE international conference on multimedia and expo (ICME '03)* (Vol. 2, pp. II–281).

Yu, D.Y.D., Deng, L.D.L., Droppo, J., Wu, J.W.J., Gong, Y.G.Y., & Acero, A. (2008). Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor. *IEEE Audio, Speech and Language Processing*, *16*(5), 1061–1070. https://doi.org/10.1109/TASL.2008.921761.

Zbilut, J.P., & Webber, C.L.J. (2006). Recurrence quantification analysis. In Akay, M. (Ed.) *Wiley encyclopedia of biomedical engineering*. Hoboken: Wiley.

Zhang, T., & Kuo, C.C.J. (1998). Hierarchical system for content-based audio classification and retrieval. In *Photonics East (ISAM, VVDC, IEMB)* (pp. 398–409). International Society for Optics and Photonics.

Zhang, H., McLoughlin, I., & Song, Y. (2015). Robust sound event recognition using convolutional neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 559–563). IEEE.