Noname manuscript No.
(will be inserted by the editor)

Relevance-based Quantization of Scattering Features for Unsupervised Mining of Environmental Audio

Vincent Lostanlen · Grégoire Lafay · Joakim Andén · Mathieu Lagrange.

Received: date / Accepted: date

Abstract The emerging field of computational ecoacoustics aims at retrieving high-level information from acoustic scenes recorded by some network of sensors. These networks gather large amounts of data requiring analysis. To decide which parts to inspect manually, we need tools that automatically mine the data, identifying recurring patterns and isolated events. This requires a similarity measure for acoustic scenes that does not impose strong assumptions on the data.

The state of the art in audio similarity measurement is the "bag-of-frames" approach, which models a recording using summary statistics of short-term audio descriptors, such as mel-frequency cepstral coefficients (MFCCs). They successfully characterise static scenes with little variability in auditory content, but cannot accurately capture scenes with a few salient events superimposed over static background. To overcome this issue, we propose a two-scale representation which describes a recording using clusters of scattering coefficients. The scattering coefficients capture short-scale structure, while the cluster model captures longer time scales, allowing for more accurate characterization of sparse events. Evaluation within the acoustic scene similarity framework demonstrates the superiority of the proposed approach.

Keywords unsupervised learning · data mining · acoustic signal processing · wavelet transforms · audio databases · content-based retrieval · nearest neighbor searches · acoustic sensors · environmental sensors.

Vincent Lostanlen
E-mail: vincent.lostanlen@nyu.edu

Grégoire Lafay
E-mail: lafaygregoire@gmail.com

Joakim Andén
E-mail: janden@flatironinstitute.org

Mathieu Lagrange
E-mail: mathieu.lagrange@cnrs.fr

1 Introduction

The amount of audio data recorded from our sonic environment has grown considerably over the past decades. In order to measure the effect of human activity and climate change on animal biodiversity, researchers have recently undertaken a massive deployment of acoustic sensors throughout the world [1–3]. In addition, recent work has explored acoustic monitoring for characterization of human pleasantness in urban areas [4, 5], as well as the prediction of annoyance due to traffic [6]. Since they bear a strong societal impact and raise many scientific challenges, we believe that these applications are of considerable interest for to signal processing community.

These questions are explored in an emerging field of computational bioacoustics, known as ecoacoustics, which is the interdisciplinary study of relationships between natural or anthropogenic sounds and their environments at the intersection of ecology, acoustics and computer science [7]. An important problem in this field is that manually analysing the recorded data to identify the quantities of interest is very costly [8]. Some sort of pre-screening is therefore required to reduce the need for human expert listening and annotation.

The most straightforward approach is to specify a closed set of sound classes, such as sounds of animals expected to live near the acoustic sensors. Computational models are trained then for these classes which are used to automatically annotate recordings [9]. A given time interval (e.g., a single day) is then represented by the number of events detected during that interval for each class. This allows the scientist to drastically reduce the amount of information requiring manual processing. However, this approach has two drawbacks. First, it relies on trained models whose behavior on unseen data (such as different sensors) is prone to errors and cannot be trusted. Second, and more importantly, it is based on *a priori* knowledge and thus cannot be considered for exploratory analysis where quantities of interest have yet to be defined.

To identify which parts need human inspection, we need tools that can detect both recurring patterns and sparsely distributed events. Identifying recurring patterns allows the user to focus on certain time points for manual annotation, while detection of more rare structures enables discovery of unforeseen phenomena.

With this aim, we need to design an algorithm for acoustic similarity retrieval, where the audio fragments judged "most similar" to a given query recording must be extracted from some larger dataset. This requires us to represent an audio recording in a way that captures its distinctive qualities. This has previously been attempted using bag-of-frames approaches [10], which describes an auditory scene recording using summary statistics of short-time features. Unfortunately, this only captures the average structure of the scene, so the approach often fails when presented with highly dynamic scenes or those characterised by a few distinct sound events. Indeed, experiments in cognitive psychology [11] and cognitive neuroscience [12] suggest that this more closely matches human acoustic perception. We believe that the fail-

ure to model such distinct events is one of the reasons why this approach is insufficient for moderate-sized evaluation datasets [13].

Solving the acoustic similarity retrieval first requires the ability to capture meaningful signal structure at small time scales. This is often achieved using mel-frequency cepstral coefficients (MFCCs). Originally developed for speech processing [14], MFCCs have recently found wider use in music information retrieval [15] and environmental audio processing [10]. A richer representation, the scattering transform, has enjoyed significant success in various audio [16] and biomedical [17] signal classification tasks. Its structure is that of a convolutional neural network [18–21], but with fixed filters. Specifically, it alternates convolutions with wavelet filters and pointwise nonlinearities to ensure time-shift invariance and time-warping stability [22].

For our task, one advantage of the scattering transform is not requiring a training step, allowing for a wider range of applications compared to learned features. Indeed, for data mining of previously unheard datasets, the properties of relevant audio structures remains to be defined, leading to an unsupervised setting.

In this work, we propose a new model for acoustic scenes, where the signal is represented at sub-second scales by scattering transforms, while larger scales are captured by an cluster model. This unsupervised model quantizes the scattering coefficients into a given number of clusters. These clusters to define a set of distances for acoustic similarity retrieval. Evaluating this approach on a scene retrieval task, we obtain significant improvements over traditional bag-of-frames and summary statistics models applied both to MFCCs and scattering coefficients.

Motivations of the proposed approach and a brief review of the state of the art in acoustic scene modeling are given in Section 2. We describe the scattering transform in Section 3, discuss feature post-processing in Section 4 and propose a cluster-based scene description in Section 5. Section 6 describes several experiments for the acoustic scene similarity retrieval task. Results are reported in Section 7.

2 Background

Computational bioacoustics refers to the numerical investigation of sound production, dispersion and reception in animals, including humans. A recent paradigm processes the audio stream in a holistic way over large time scales, without assuming that a single species is present throughout the recording. Automated systems within this paradigm attempt to infer global properties of bioacoustic scenes, including biodiversity indices [23] and migration patterns [24]. They also mark time intervals of particular interest for detailed human inspection [25].

A closely related field is that of urban sound environment analysis. In this context, a popular approach is the bag of frames (BOF), first applied to the problem by Aucouturier et al. [10]. It models an auditory scene using high-level

summary statistics computed from local features, typically implemented by Gaussian mixture models (GMMs) of MFCCs. Recently, this representation was shown to perform comparably to direct averaging of the features for a variety of tasks [13]. This contrasts with the typical morphology of acoustic scenes, a "skeleton of events on a bed of textures," where a few discrete sound events are superimposed upon a stationary acoustic background [26]. Such events are not well-characterised by summarizing short-term features, but are better described by large-scale temporal evolution of auditory scenes. The latter approach should therefore prove more fruitful in measuring auditory scene similarity.

This statement has some support in auditory psychology as well as sound synthesis based on summary statistics [27]. Studies in the cognitive psychology of urban sound environments have shown that global sound level (perceived or measured) is not sufficient to fully characterise an acoustic scene [4, 28]. Instead, cognitive processes such as sound environment quality perception [11] or loudness judgment [29] seem to rely upon higher-level cognitive attributes. These typically include the identities of the sound sources which constitute the scene. It has been shown that, if available, the complete description of the scene in terms of event occurrences is powerful enough to reliably predict high-level cognitive classes. For example, in urban areas the presence of birds is likely to be heard in parks and are therefore strong pleasantness indicators. Consequently, research in sound perception is now strongly focused on the contribution of specific sound sources in the assessment of sound environments [5, 30]. Although the complete set of events occurring within a given auditory stream may not be discernable even to human expects, research has shown that a small set of events (so-called markers) suffice to reliably predict many high-level attributes.

From a cognitive psychology perspective, the consensus is therefore that only a few distinct events are sufficient to describe an auditory scene, in contrast to BOF models which treat each observation separately and do not capture their temporal structure. A method that takes this knowledge into account could therefore have potential for great impact in acoustic scene modeling, given a rich enough representation of these distinct events.

3 Wavelet scattering

Local invariance to time-shifting and stability to time-warping are necessary when representing acoustic scenes for similarity measurement. The scattering transform is designed to satisfy these properties while retaining high discriminative power. It is computed by applying auditory and modulation wavelet filter banks alternated with complex modulus nonlinearities.

3.1 Invariance and stability in audio signals

The notion of invariance to time-shifting plays an essential role in acoustic scene similarity retrieval. Indeed, recordings may be shifted locally in time without affecting similarity to other recordings. To discard this superfluous source of variability, signals are first mapped into a time-shift invariant feature space. These features are then used to calculate similarities. Since the features ensure invariance, it does not have to be learned when constructing the desired similarity measure.

Formally, given a signal $\boldsymbol{x}(t)$, we would like its translation $\boldsymbol{x_c}(t) = \boldsymbol{x}(t-c)$ to be mapped to the same feature vector provided that $|c| \ll T$ for some maximum duration T that specifies the extent of the time-shifting invariance. We can also define more complicated transformations by letting c vary with t. In this case, we have $\boldsymbol{x_\tau}(t) = \boldsymbol{x}(t - \tau(t))$ for some function τ, which performs a time-warping of $\boldsymbol{x}(t)$ to obtain $\boldsymbol{x_\tau}(t)$. Time-warpings model various changes, such as small variations in pitch, reverberation, and rhythmic organization of events. These make up an important part of intra-class variability among natural sounds, so representations must be robust with respect to such transformations.

The wavelet scattering transform, described below, has both of these desired properties: invariance to time-shifting and stability to time-warping. The stability condition can be formulated as a Lipschitz continuity property, which guarantees that the feature transforms of $\boldsymbol{x}(t)$ and $\boldsymbol{x_\tau}(t)$ are close together if $|\tau'(t)|$ is bounded by a small constant.

3.2 Wavelet scalogram

Our convention for the Fourier transform of a continuous-time signal $\boldsymbol{x}(t)$ is $\hat{\boldsymbol{x}}(\omega) = \int_{-\infty}^{+\infty} x(t) \exp(-\mathrm{i}2\pi\omega t)\ \mathrm{d}t$. Let $\boldsymbol{\psi}(t)$ a complex-valued analytic bandpass filter of central frequency ξ_1 and bandwidth ξ_1/Q_1, where Q_1 is the quality factor of the filter. A filter bank of wavelets is built by dilating $\boldsymbol{\psi}(t)$ according to a geometric sequence of scales $2^{\gamma_1/Q_1}$, obtaining

$$\boldsymbol{\psi_{\gamma_1}}(t) = 2^{-\gamma_1/Q_1} \boldsymbol{\psi}(2^{-\gamma_1/Q_1} t). \tag{1}$$

The variable γ_1 is a scale (an inverse log-frequency) taking integer values between 0 and $(J_1 Q_1 - 1)$, where J_1 is the number of octaves spanned by the filter bank. For each γ_1, the wavelet $\boldsymbol{\psi_{\gamma_1}}(t)$ has a central frequency of $2^{-\gamma_1/Q_1}\xi_1$ and a bandwidth of $2^{-\gamma_1/Q_1}\xi_1/Q_1$ resulting in the same quality factor Q_1 as $\boldsymbol{\psi}$. In the following, we set ξ_1 to 20 kHz, J_1 to 10, and the quality factor Q_1, which is also the number of wavelets per octave, to 8. This results in the wavelet filters covering the whole range of human hearing, from 20 Hz to 20 kHz. Setting $Q_1 = 8$ results in filters whose bandwidth approximates an equivalent rectangular bandwidth (ERB) scale [31].

The wavelet transform of an audio signal $\boldsymbol{x}(t)$ is obtained by convolution with all wavelet filters. Applying a pointwise complex modulus the transform

yields the wavelet scalogram

$$\boldsymbol{x_1}(t, \gamma_1) = |\boldsymbol{x} * \boldsymbol{\psi_{\gamma_1}}|(t). \tag{2}$$

The scalogram bears resemblance to the constant-Q transform (CQT), which is derived from the short-term Fourier transform (STFT) by averaging the frequency axis into constant-Q subbands of central frequencies $2^{-\gamma_1/Q_1}\xi_1$. Indeed, both time-frequency representations are indexed by time t and log-frequency γ_1. However, contrary to the CQT, the scalogram reaches a better time-frequency localization across the whole frequency range, whereas the temporal resolution of the traditional CQT is fixed by the support of the STFT analyzing window. Therefore, the scalogram has a better temporal localization at high frequencies than the CQT, at the expense of a greater computational cost since the inverse fast Fourier transform routine must be called for each wavelet $\boldsymbol{\psi_{\gamma_1}}$ in the filter bank. However, this allows us to observe amplitude modulations at fine temporal scales in the scalogram, down to $2Q_1/\xi_1$ for $\gamma_1 = 0$, of the order of 1 ms given the aforementioned values of Q_1 and ξ_1.

To obtain the desired invariance and stability properties, the scalogram is averaged in time using a lowpass filter $\boldsymbol{\phi}(t)$ with cut-off frequency $1/T$ (and approximate duration T), to get

$$\mathbf{S_1}\boldsymbol{x}(t, \gamma_1) = \boldsymbol{x_1}(\cdot, \gamma_1) * \boldsymbol{\phi}(t), \tag{3}$$

which is known as the set of first-order scattering coefficients. They capture the average spectral envelope of $\boldsymbol{x}(t)$ over scales of duration T and where the spectral resolution varying with constant Q. In this way, they are closely related to the mel-frequency spectrogram and related features, such as MFCCs.

3.3 Extracting modulations with second-order scattering

In auditory scenes, short-time amplitude modulations may be caused by a variety of rapid mechanical interactions, including collision, friction, turbulent flow, and so on. At longer time-scales, they also account for higher-level attributes of sound, such as prosody in speech or rhythm in music. Although they are discarded while filtering $\boldsymbol{x_1}(t, \gamma_1)$ into the time-shift invariant representation $\mathbf{S_1}\boldsymbol{x}(t, \gamma_1)$, they can be recovered from $\boldsymbol{x_1}(t, \gamma_1)$ by a second wavelet transform and another complex modulus.

We define second-order wavelets $\boldsymbol{\psi_{\gamma_2}}(t)$ in the same way as the first-order wavelets, but with parameters ξ_2, J_2, and Q_2. Consequently, they have central frequencies $2^{-\gamma_2/Q_2}\xi_2$ for γ_2 taking values between 0 and (J_2Q_2-1). While this abuses notation slightly, the identity of the wavelets should be clear from context. The amplitude modulation spectrum resulting from a wavelet modulus decomposition using these second-order wavelets is then

$$\boldsymbol{x_2}(t, \gamma_1, \gamma_2) = |\boldsymbol{x_1} * \boldsymbol{\psi_{\gamma_2}}|(t, \gamma_1). \tag{4}$$

In the following, we set ξ_2 to 2.5 kHz, Q_2 to 1, and J_2 to 12. Lastly, the low-pass filter $\boldsymbol{\phi}(t)$ is applied to $\boldsymbol{x_2}(t, \gamma_1, \gamma_2)$ to guarantee local invariance to time-shifting, which yields the second-order scattering coefficients

$$\mathbf{S_2}\boldsymbol{x}(t, \gamma_1, \gamma_2) = (\boldsymbol{x_2}(\cdot, \gamma_1, \gamma_2) * \boldsymbol{\phi})(t). \tag{5}$$

The scattering transform $\mathbf{S}\boldsymbol{x}(t, \gamma)$ consists of the concatenation of first-order coefficients $\mathbf{S_1}\boldsymbol{x}(t, \gamma_1)$ and second-order coefficients $\mathbf{S_1}\boldsymbol{x}(t, \gamma_1, \gamma_2)$ into a feature matrix $\mathbf{S}\boldsymbol{x}(t, \gamma)$, where γ denotes either γ_1 or (γ_1, γ_2). While higher-order scattering coefficients can be calculated, for the purposes of our current work, the first and second order are sufficient. Indeed, higher-order scattering coefficients have been shown to contain reduced energy and are therefore of limited use [32].

3.4 Gammatone wavelets

[Fig. 1 about here.]

Wavelets $\boldsymbol{\psi_{\gamma_1}}(t)$ and $\boldsymbol{\psi_{\gamma_2}}(t)$ are designed as fourth-order Gammatone wavelets with one vanishing moment [33], and are shown in Figure 1. In the context of auditory scene analysis, the asymmetric envelopes of Gammatone wavelets are more biologically plausible than the symmetric, Gaussian envelopes of the more widely used Morlet wavelets. Indeed, it allows to reproduce two important psychoacoustic effects in the mammalian cochlea: the asymmetry of temporal masking and the asymmetry of spectral masking [31]. The asymmetry of temporal masking is the fact that a masking noise has to be louder if placed after the onset of a stimulus rather than before. Likewise, because critical bands are skewed towards higher frequencies, a masker tone has to be louder if it is above the stimulus in frequency rather than below.

It should also be noted that Gammatone wavelets follow the typical amplitude profile of natural sounds, beginning with a relatively sharp attack and ending with a slower decay. As such, they are similar to filters discovered automatically by unsupervised encoding of natural sounds [34]. This suggests that, despite being hand-crafted and not learned, Gammatone wavelets provide a sparser time-frequency representation of acoustic scenes compared to other variants.

4 Feature design

Before constructing models for similarity estimation, it is beneficial to process scattering coefficients to improve invariance, normality, and generalization power. In this section, we review two transformations which achieve these properties: logarithmic compression and standardisation.

4.1 Logarithmic compression

Many algorithms in pattern recognition, including nearest neighbor classifiers and SVMs, tend to work best when all features follow a standard normal distribution across all training instances [35]. Yet the distribution of the scattering coefficients is skewed towards larger values. We can reduce this skewness by applying a pointwise concave transformation to all coefficients. In particular, we find that the logarithm performs particularly well in this respect. Figure 2 shows the distribution of an arbitrarily chosen scattering coefficient over the DCASE 2013 dataset, before and after logarithmic compression.

[Fig. 2 about here.]

Taking the logarithm of a magnitude spectrum is ubiquitous in audio signal processing. Indeed, it is corroborated by the Weber-Fechner law in psychoacoustics, which states that the sensation of loudness is roughly proportional to the logarithm of the acoustic pressure. We must also recall that the measured amplitude of sound sources often decays polynomially with the distance to the microphone–a source of spurious variability in scene classification. Logarithmic compression linearise this dependency, facilitating the construction of powerful invariants at the classifier stage.

4.2 Standardisation

Let $\mathbf{S}\boldsymbol{x}(\gamma, n)$ be a dataset, where γ and n denote feature and sample indices, respectively. Many algorithms operate better on features which have zero mean and unit variance to avoid mismatch in numeric ranges [35]. To standardise $\mathbf{S}\boldsymbol{x}(\gamma, n)$, we subtract the sample mean vector $\mu[\mathbf{S}\boldsymbol{x}(\gamma)]$ from $\mathbf{S}\boldsymbol{x}(\gamma, n)$ and divide the result by the sample standard deviation vector $\sigma[\mathbf{S}\boldsymbol{x}](\gamma)$. The vectors $\mu[\mathbf{S}\boldsymbol{x}(\gamma)]$ and $\sigma[\mathbf{S}\boldsymbol{x}](\gamma)$ are estimated from the entire dataset.

5 Acoustic scene similarity retrieval

As discussed in Section 2, results in sound perception suggest the appropriateness of source-driven representations of auditory scenes for predicting high-level properties. While this can be addressed in the supervised case using late integration of discriminative classifiers [16], this is not directly feasible in the unsupervised case. As the detection of events is still an open problem [36], we consider in this paper a generic quantization scheme in order to identify and represent time intervals of the scene that are coherent, thus likely to be dominated by a given source of interest.

Given a set of d-dimensional feature vectors $X_u = \{x_1^u, \ldots, x_L^u\}$, extracted from the scene s_u, where $u = \{1, 2, \ldots, U\}$, we would like to partition X_u into a set $C_u = \{c_1^u, \ldots, c_M^u\}$ of M clusters. This partition is obtained by minimizing the variance of each cluster and known as a k-means clustering

[37]. Each scene s_u is then described by a set of clusters C_u. Note that this quantization approach differs from unsupervised learning schemes such as the ones studied in [38], where the scene features are projected in a dictionary learned from the entire dataset. Here, with the aim of better balancing the influence of salient sound events and texture-like sounds on the final decision, the similarity between two scenes is computed based on the similarity of their centroids.

The similarity between the scene centroids μ_m^u over the entire dataset, is computed using a radial basis function (RBF) kernel K combined with a local scaling method [39]:

$$K_{mn}^{uv} = \exp\left(-\frac{\|\mu_m^u - \mu_n^v\|^2}{\|\mu_m^u - \mu_{m,q}^u\|\|\mu_n^v - \mu_{n,q}^v\|}\right). \tag{6}$$

Here, $\mu_{m,q}^u$ and $\mu_{n,q}^v$ are the q^{th} nearest neighbors to the centroids μ_m^u and μ_n^v, respectively, and $\|\cdot\|$ denotes the Euclidean norm.

To compute the similarity between two scenes, we consider several centroid-based similarity metrics:

- Relevance-based Quantization closest similarity (*RbQ-c*): the similarity between two scenes s_u and s_v is equal to the largest similarity between their centroids
$$\max_{m,n} K_{mn}^{uv}, \tag{7}$$
- Relevance-based Quantization average similarity (*RbQ-a*): the similarity between two scenes s_u and s_v is equal to the average of their centroid similarities
$$\frac{1}{M^2}\sum_{m,n} K_{mn}^{uv} \tag{8}$$
and,
- Relevance-based Quantization weighted similarity (*RbQ-w*): the similarity between two scenes is computed using a variant of the earth mover's distance applied to the set of centroids each weighted by the number of frames assigned to its cluster.

For *RbQ-w*, each scene is represented by a signature

$$p_u = \{(\mu_1^u, w_1^u), (\mu_2^u, w_2^u), \ldots, (\mu_M^u, w_M^u)\},$$

where each of the M centroids $\mu_1^u, \ldots, \mu_M^u$ are paired with corresponding weights $w_1^u, \ldots, w_M^u$. The weight w_m^u for the mth centroid μ_m^u is the number of frames belonging to a particular cluster. The similarity between scenes is then given by a cross-bin histogram distance known as the non-normalized earth mover's distance $\widehat{\text{EMD}}$ introduced by [40]. The $\widehat{\text{EMD}}$ computes the distance between two histograms by finding the minimal cost for transforming one histogram into the other, where cost is measured by the number of transported histogram counts multiplied by the a dissimilarity measure between the histogram bins. Here, that measure is given by $1 - K_{mn}^{uv}$.

6 Experiments

To evaluate the representations introduced in the previous section, we apply it to the acoustic scene similarity retrieval task. Results demonstrate the improved performance of the relevance-based quantization of scattering coefficients compared to baseline methods using summary statistics of MFCCs. The implementations of the presented methods and the experimental protocol are available online.[1]

6.1 Dataset

The experiments in this paper are carried out on the DCASE 2013 dataset [36]. This dataset consists of two parts, a public and a private subset, each made up of 100 acoustic scene recordings sampled at 44100 Hz and 30 seconds in duration. The dataset is evenly divided into 10 acoustic scene classes. The recordings were made by three different recordists at a wide variety of locations in the Greater London area over a period of several months. No systematic variations in the recordings covaried with scene type: all recordings were made under moderate weather conditions, at varying times of day and week, and each recordist recorded each scene type. As a result, the dataset enjoys significant intra-class diversity while remaining of manageable size, making it suitable for evaluation of algorithmic design choices [13].

6.2 Feature design

We perform our experiments using both scattering coefficients and MFCCs. For the scattering transform, each 30-second scene is described by 128 vectors of dimension 1367 computed with half-overlapping windows $\phi(t)$ of duration $T = 372$ ms, for a total of 24 s. Here we discard 3 seconds from the beginning and end of the scene to avoid boundary artifacts. We also conduct experiments with and without logarithmic compression of the scattering coefficients (see Section 4.1).

MFCCs are computed for windows of 50 ms and hops of 25 ms with full frequency range. The standard configuration of 39 coefficients coupled with an average-energy measure performs best in preliminary tests, so we use this in the following. We average the coefficients using 250 ms long non-overlapping windows so that each window represents structures of scales close to that of scattering coefficients.

6.3 Algorithm

The evaluation is performed on the private part of the DCASE 2013 dataset. As a metric, we use the precision at rank k ($p@k$). This number is computed

[1] `https://github.com/mathieulagrange/paperRelevanceBasedSimilarity`

by taking a query item and counting the number of items of the same class within the k closest neighbors, and then averaging over all query items. We determine these neighbors using one of the proposed similarity measures *RbQ-c*, *RbQ-a*, or *RbQ-w*. We compute $p@k$ for $k = \{1, \ldots, 9\}$, since each class only has 10 items. Note that $p@1$ is equal to the classification accuracy obtained by the nearest-neighbor classifier in a leave-one-out cross-validation setting.

The *RbQ* measures are compared to commonly used early integration approach *early*, which consists in averaging over time the feature set of each scene, resulting in one feature vector per scene. The distance on this average feature vector is then used to determine $p@k$. For the BOF approach of Aucouturier et al. [10], GMMs are estimated for each scene using the expectation-maximization (EM) algorithm [41, 42]. The similarity between a given pair of scene GMMs is then calculated through Monte Carlo sampling approximation. To ensure convergence of the EM algorithm for the scattering features, we reduce their dimension from 1367 to 30 by projecting the features onto the top 30 principal components of the dataset. the number of Gaussians is optimized for each type of features by grid search in the range $[2, 20]$. Best $p@5$ is reached with 8 and 4 Gaussians, respectively for MFCCs and scattering features. Recommended number of Gaussians for MFCCs given in [10] is 10.

The scaling parameter q of the RBF kernels (see Eq. 6) is set to 10% of the number of data points to cluster. For each method, the numbers of clusters is set to 8.

7 Results

Results for the acoustic scene similarity retrieval task demonstrate the superiority of the logarithmically compressed scattering features compared to MFCCs. Combining these with the *RbQ* cluster model, we obtain significant improvements over traditional BOF and summary statistic measures.

Baselines

As seen in Figure 3, the BOF approach behaves similarly to the early approach when applied both to the MFCCs and the scattering features.

The early approach being simpler in terms of implementation and runtime complexity, we retain this method as baseline for the remainder of the experiments.

[Fig. 3 about here.]

Logarithmic compression

Figure 4 shows that logarithmic compression of the scattering features is beneficial. For clarity sake, data is shown for the *early* approach only, but a equivalent gain is achieved for the relevance-based quantization approaches.

[Fig. 4 about here.]

MFCC vs. scattering transform

Irrespective of the rank k considered, best result is achieved for the scattering transform with logarithmic compression using the *RbQ-c* approach. Overall, log-compressed scattering coefficients systematically outperform MFCCs. This is to be expected since the scattering coefficients capture larger-scale modulations, as opposed to MFCCs which only describe the short-time spectral envelope.

Relevance-based quantization vs. early integration

For the scattering transform, both *RbQ-c* and *RbQ-w* outperform *early*, thus confirming the benefits of using an relevance-based quantization (*RbQ*) to improve the similarity measures between the scenes. However, it is worth noting that *RbQ-a* performs comparably to or worse than *early*, showing that the discriminant information is destroyed by averaging the contributions from all centroids. This result is in line with the findings of [13]. To take advantage of such a representation, we need to select certain representative centroids when comparing quantized objects.

Furthermore, it appears that *RbQ-c* is better able to characterise the classes compared to *RbQ-w*. Although not the only way of incorporating the number of frames associated to each centroid, the earth mover's distance is a rather natural of doing so. Its worse performance therefore suggests that including this information may not always be desirable. Indeed, nothing a priori indicates that the discriminant information between two scenes lays within the majority of their frames. On the contrary, two similar environments may share a lot of similar sound sources with only a few sources discriminating between them.

With p@5 as our metric (cf. [10] and [13]), we see that replacing MFCCs by the logarithmically compressed scattering transform increases performance from 0.31 to 0.49. In addition, the relevance-based quantization using the closest similarity (*RbQ-c*) further improves the performance to 0.54 for a global increase of 0.23. Vincent : It is worth reminding what the baselines are.

[Fig. 5 about here.]

8 Conclusion

This paper presents a new approach for modeling acoustic scenes based on scattering transforms at small scales and cluster-based representations at large scales. Compared to traditional BOF and summary statistics models, this representation allows for the characterization of distinct sound events superimposed on a stationary texture, a concept which has strong grounding in the cognitive psychology literature. To adequately capture such distinct events, we

develop a cluster-based model and validate it using experiments on acoustic scene similarity retrieval. For this task, we show significant improvements over the traditional BOF and summary statistics models based on both standard MFCCs and scattering features. These outcomes shall be studied further in future work by considering larger databases and emerging tasks in computational bioacoustics [8].

Availability of data and materials

The dataset supporting the conclusions of this article is available in the dcase2013 repository, `http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/description.html`. The software supporting the conclusions of this article is available in

- Project name: article2017EstimationAmbiance
- Project home page: `https://github.com/mathieulagrange/paperRelevanceBasedSimilarity`
- Programming language: Matlab
- License: GNU GPL
- Any restrictions to use by non-academics: license needed

Competing interests

The authors declare that they have no competing interests.

Funding

This study is co-funded by the ANR under project reference ANR-16-CE22-0012.

Authors' contributions

GL and VL carried out the numerical experiments and drafted the manuscript. VL, GL, JA and ML participated in the design of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

References

[1] P. S. Warren et al. "Urban bioacoustics: It's not just noise". In: *Animal behaviour* 71.3 (2006), pp. 491–502.

[2] S. R. Ness et al. "The Orchive: Data mining a massive bioacoustic archive." In: *International Workshop on Machine Learning for Bioacoustics* (2013).

[3] D. Stowell and M. D. Plumbley. “Large-scale analysis of frequency modulation in birdsong databases”. In: *Methods in Ecology and Evolution* 11 (2013).

[4] F. Guyot et al. “Urban sound environment quality through a physical and perceptive classification of sound sources: A cross-cultural study”. In: *Proceedings Forum Acusticum, Budapest, Hungary*. 2005.

[5] P. Ricciardi et al. “Sound quality indicators for urban places in Paris cross-validated by Milan data”. In: *The Journal of the Acoustical Society of America* 138.4 (2015), pp. 2337–2348.

[6] J. R. Gloaguen et al. “Estimating traffic noise levels using acoustic monitoring: a preliminary study”. In: *Workshop on Detection and Classification of Acoustic Scenes and Events*. 2016.

[7] B. C. Pijanowski et al. “What is soundscape ecology? An introduction and overview of an emerging new science”. In: *Landscape Ecology* 26.9 (2011), pp. 1213–1232.

[8] J. Wimmer et al. “Sampling environmental acoustic recordings to determine bird species richness”. In: *Ecological Applications* 23.6 (2013), pp. 1419–1428.

[9] L. Zhang et al. “Classifying and ranking audio clips to support bird species richness surveys”. In: *Ecological Informatics* 34 (2016), pp. 108–116. DOI: 10.1016/j.ecoinf.2016.05.005. URL: https://eprints.qut.edu.au/96243/.

[10] J.-J. Aucouturier, B. Defreville, and F. Pachet. “The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music”. In: *The Journal of the Acoustical Society of America* 122.2 (2007), pp. 881–891.

[11] D. Dubois, C. Guastavino, and M. Raimbault. “A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories”. In: *Acta Acustica United with Acustica* 92.6 (2006), pp. 865–874.

[12] I. Nelken. “Processing of complex stimuli and natural scenes in the auditory cortex”. In: *Current Opinion in Neurobiology* 14.4 (2004), pp. 474–480.

[13] M. Lagrange et al. “The bag-of-frames approach: A not so sufficient model for urban soundscapes”. In: *JASA Express Letters* 138.5 (Oct. 2015), pp. 487–492.

[14] S. Davis and P. Mermelstein. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (1980), pp. 357–366.

[15] B. Logan. “Mel frequency cepstral coefficients for music modeling”. In: *Proceedings of the International Symposium on Music Information Retrieval*. 2000.

[16] J. Andén and S. Mallat. “Deep scattering spectrum”. In: *IEEE Transactions on Signal Processing* 62.16 (2014), pp. 4114–4128.

[17] V Chudáček et al. “Scattering transform for intrapartum fetal heart rate variability fractal analysis: A case study”. In: *IEEE Transactions on Biomedical Engineering* (2013).

[18] H. Lee et al. “Unsupervised feature learning for audio classification using convolutional deep belief networks”. In: *Proc. NIPS*. 2009.

[19] V. Lostanlen and C.-E. Cella. “Deep convolutional networks in the pitch spiral for music instrument classification”. In: *Proceedings of the International Society for Music Information Retrieval Conference*. ISMIR. 2016.

[20] Y. Aytar, C. Vondrick, and A. Torralba. “SoundNet: Learning sound representations from unlabeled video”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 892–900.

[21] R. Arandjelović and A. Zisserman. “Look, Listen and Learn”. In: (2017). eprint: 1705.08168.

[22] S. Mallat. “Group Invariant Scattering”. In: *Communications on Pure and Applied Mathematics* 65.10 (2012), pp. 1331–1398. ISSN: 00103640. DOI: 10.1002/cpa.21413. arXiv: 1101.2286.

[23] R. Bardeli et al. “Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring”. In: *Pattern Recognition Letters* 31.12 (2010), pp. 1524–1534.

[24] M. K. Obrist et al. “Bioacoustics approaches in biodiversity inventories”. In: *Abc Taxa* 8 (2010), pp. 68–99.

[25] S. S. Rosenstock et al. “Landbird counting techniques: Current practices and an alternative”. In: *The Auk* 119.1 (2002), pp. 46–53.

[26] I. Nelken and A. de Cheveigné. “An ear for statistics”. In: *Nature Neuroscience* 16.4 (2013), pp. 381–382.

[27] J. H. McDermott, M. Schemitsch, and E. P. Simoncelli. “Summary statistics in auditory perception”. In: *Nature neuroscience* 16.4 (2013), pp. 493–498.

[28] J. Kang. *Urban sound environment*. CRC Press, 2006.

[29] S. Kuwano et al. “Memory of the loudness of sounds in relation to overall impression”. In: *Acoustics Science and Technics* 4.24 (2003).

[30] C. Lavandier and B. Defréville. “The contribution of sound source characteristics in the assessment of urban soundscapes”. In: *Acta Acustica United with Acustica* 92.6 (2006), pp. 912–921.

[31] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and models*. Vol. 22. Springer Science & Business Media, 2013.

[32] I. Waldspurger. “Exponential decay of scattering coefficients”. In: *Proc. SampTA*. 2017, pp. 143–146. DOI: 10.1109/SAMPTA.2017.8024473.

[33] A. Venkitaraman, A. Adiga, and C. S. Seelamantula. “Auditory-motivated Gammatone wavelet transform”. In: *Signal Processing* 94 (2014), pp. 608–619.

[34] E. C. Smith and M. S. Lewicki. “Efficient auditory coding”. In: *Nature* 439.7079 (2006), pp. 978–982.

[35] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. *A practical guide to support vector classification*. Tech. rep. National Taiwan University, Taiwan, 2003.

[36] D. Stowell et al. "Detection and classification of acoustic scenes and events". In: *IEEE Transactions on Multimedia* 17.10 (2015), pp. 1733–1746.

[37] S. Lloyd. "Least squares quantization in PCM". In: *IEEE Trans. Inf. Theory* 28.2 (1982), pp. 129–137. ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056489.

[38] V. Bisot et al. "Acoustic scene classification with matrix factorization for unsupervised feature learning". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 6445–6449.

[39] L. Zelnik-Manor and P. Perona. "Self-tuning spectral clustering". In: *Advances in Neural Information Processing Systems. (NIPS) No. 17*. MIT Press, Cambridge, MA. 2004, pp. 1601–1608.

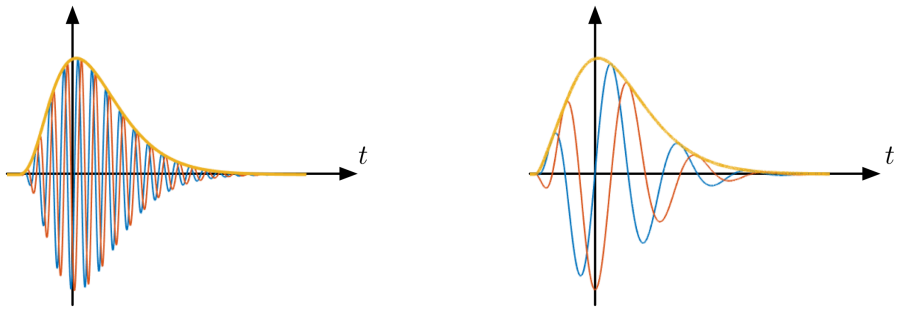
[40] O. Pele and M. Werman. "A linear time histogram metric for improved SIFT matching". In: *European Conference on Computer Vision*. Springer. 2008, pp. 495–508.

[41] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *J. Royal Stat. Soc. B Stat. Methol.* 39.1 (1977), pp. 1–38. ISSN: 00359246. URL: http://www.jstor.org/stable/2984875.

[42] T. K. Moon. "The expectation-maximization algorithm". In: *IEEE Signal processing magazine* 13.6 (1996), pp. 47–60.

List of Figures

1 Gammatone wavelets $\psi(t)$ in the time domain with quality factors (a) $Q = 4$ and (b) $Q = 1$. Oscillations (red, blue) are the real and imaginary parts. The envelope (yellow) is the complex modulus. 18
2 Histogram of values taken by the first-order scattering coefficient $\mathbf{S}\boldsymbol{x}(\gamma)$, corresponding to a central acoustic frequency of 302 Hz, (a) before and (b) after logarithmic compression. 19
3 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k ($p@k$) obtained for scattering with or without logarithmic compression, as a function of the rank k. 20
4 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k ($p@k$) obtained for scattering with or without logarithmic compression, as a function of the rank k. 21
5 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k ($p@k$) obtained for MFCCs and scattering with logarithmic compression. 22



(a) (b)

Fig. 1 Gammatone wavelets $\psi(t)$ in the time domain with quality factors (a) $Q = 4$ and (b) $Q = 1$. Oscillations (red, blue) are the real and imaginary parts. The envelope (yellow) is the complex modulus.

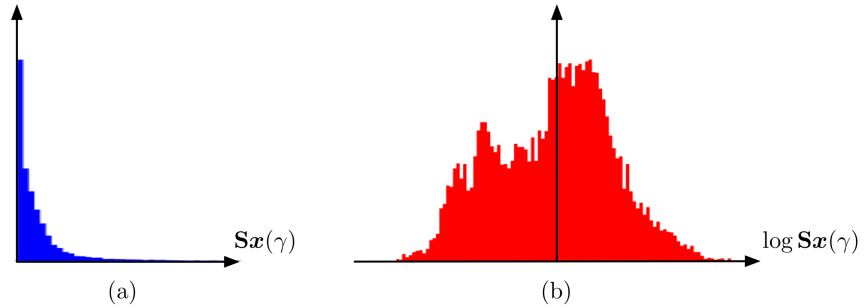


Fig. 2 Histogram of values taken by the first-order scattering coefficient $\mathbf{S}\boldsymbol{x}(\gamma)$, corresponding to a central acoustic frequency of 302 Hz, (a) before and (b) after logarithmic compression.

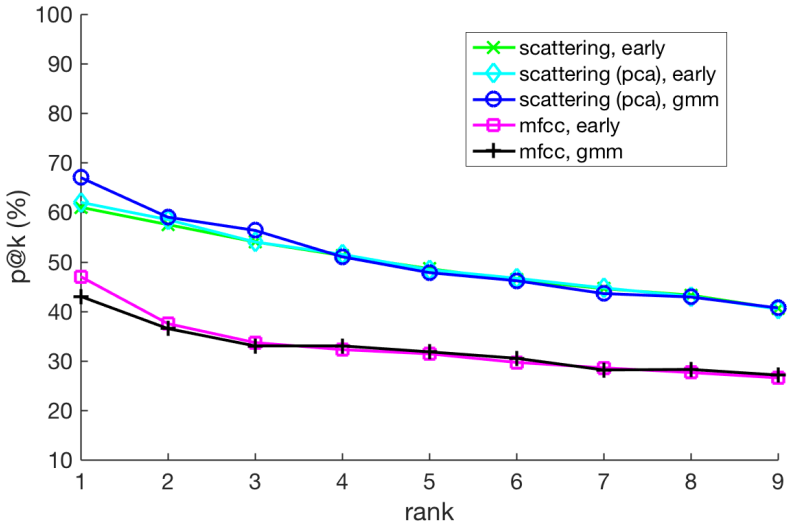


Fig. 3 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k ($p@k$) obtained for several baseline approaches, as a function of the rank k.

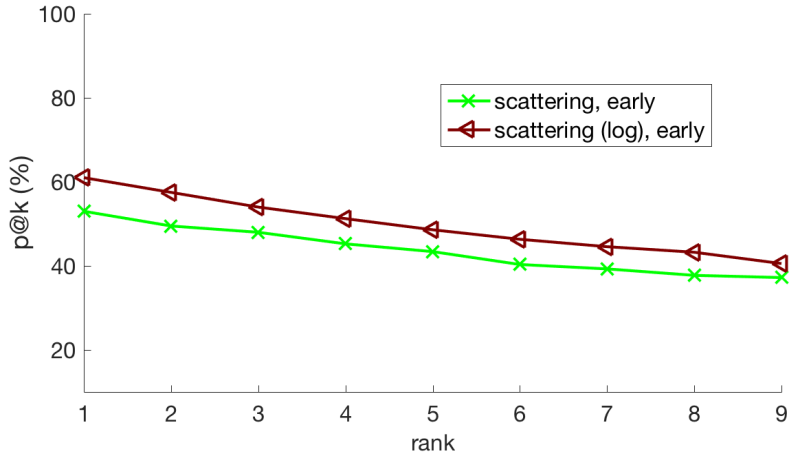


Fig. 4 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k ($p@k$) obtained for scattering with or without logarithmic compression, as a function of the rank k.

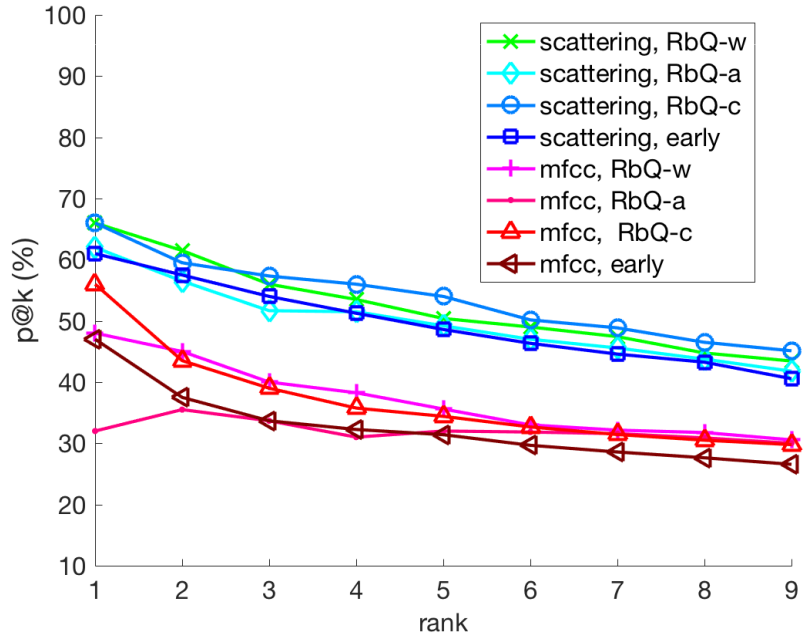


Fig. 5 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k ($p@k$) obtained for MFCCs and scattering with logarithmic compression.