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, and Mathieu Lagrange

the date of receipt and acceptance should be inserted later

Abstract The emerging field of computational ecoacoustics aims at retrieving high-level information from acoustic scenes recorded by some network of sensors. Large amounts of unlabeled data are gathered using such networks that needs to be analyzed. In order to identify which parts require human inspection, one needs tools that can mine data and identify structures such as recurring patterns or isolated events. To do so, one needs to be able to measure the level of similarity among the data without strong assumptions.

State of the art approaches for computing such a similarity is the "bag-of-frames" approach, i.e., modeling the audio signal using the summary statistics of short-term audio descriptors, such as mel-frequency cepstral coefficients (MFCCs). Although they successfully characterize static scenes with little variability in auditory content, they are unable to accurately discriminate scenes with few salient events superimposed over a textured background. To overcome this issue, we propose a two-scale representation which describes a signal at a short scale using scattering transforms while relying on a cluster model to characterize its larger-scale structure that better handle the required modeling of sparse events. Experiments performed using 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.

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

This work is supported by the ERC InvariantClass grant 320959.

Address(es) of author(s) should be given

activity and climate change on animal biodiversity, researchers have recently undertaken a massive deployment of acoustic sensors throughout the world [**warren2006urban**, **NessSST13**, **stowell13b**]. In addition, recent work has explored acoustic monitoring for characterization of human pleasantness in urban areas [**guyot2005urban**, **ricciardi2015sound**], as well as the prediction of annoyance due to traffic [**gloaguen**]. Since they bear a strong societal impact and raise many scientific challenges, we believe that these applications are of considerable interest for the signal processing community.

This work falls in the realm of ecoacoustics, an emerging field in computational bioacoustics, which is the interdisciplinary study of relationships between natural or anthropogenic sounds and their environments at the intersection of ecology, acoustics and computer science [**krause**]. An important problem in this field is that manually analysing the recorded data to identify the quantities of interest is very costly [**wimmer2013sampling**]. Some sort of pre-screening is therefore needed to reduce the need for human expert listening and annotation.

The most straightforward approach is to assume a closed set of events of interests like sounds of animals expected to live near the acoustic sensors, train models for them and perform automatic annotation []. A given time interval, for example a single day, is then represented by the number of events within the predefined typology that have been detected in that period of time. This approach allows the scientist to drastically reduce the amount of information that needs to be processed. However, this approach has two drawbacks. The first is the reliance to trained models whose behavior on unseen data (such as different sensors) is prone to errors and cannot be always be trusted. More importantly, it is based on *a priori* knowledge and thus cannot be considered for exploratory analysis where the quantities of interest have yet to be defined.

In order to identify which parts needs human inspection, one needs tools that can mine data and annotate to whether recurring patterns or sparsely distributed events to study anomalous or scarce behaviors. The latter allowing the scientist to then annotate only a representative audio of the periods of time with recurring patterns and the former to potentially discover unforseen phenomena.

With this aim, one needs to design an algorithm to perform an acoustic similarity retrieval, where the audio fragments judged “most similar” to a given query section are extracted from some larger dataset. This requires the audio fragments to be represented in some way that captures their distinctive qualities. This has been achieved by bag-of-frames approaches [**aucouturier2007bag**], which attempt to describe the auditory scene recording using summary statistics. Unfortunately, since this only captures the average structure of the scene, the approach often fails when presented with scenes which are dynamic and contain episodic, distinct sound events that discriminate the scene from others. Indeed, experiments in cognitive psychology [**dubois2006cognitive**] and cognitive neurosciences [**nelken2004processing**] suggest that this matches more closely human acoustic perception of acoustic scenes. We believe that the failure to model such distinct events is one of the reason why this approach is found to be insufficient when considering evaluation datasets of typical sizes [**lagrange:hal-01082501**].

Addressing this task first requires the ability to capture meaningful aspects within the signal at small time-scales. In other words, how do we represent an extract of the auditory scene signal over a sub-second interval? A standard tool for this are the mel-frequency cepstral coefficients (MFCCs), originally developed for speech processing [**davis-mermelstein**], but which have found wider use in recent years in music information retrieval [**logan**] and environmental audio processing [**aucouturier2007bag**]. A richer set of features is found in the scattering transform, a generic signal representation which has enjoyed significant success in various audio classification tasks [**Anden2014**, **chudacek**]. This is because the scattering transform cascades wavelet convolutions and pointwise nonlinearities, thus imitating the architecture of a convolutional neural network [**lee**, **lostanlen-deep-spiral**, **soundnet**, **arandjelovic-zisserman**].

For the task at at hand, scattering features have the advantage of not requiring training and thus potentially lead to wider range of application than trained features. Indeed, the task at hand is the mining of yet unheard audio datasets and the properties of the audio that are of interest remains to be defined, thus leading to an unserpervised setting.

We also 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 unsupervised method that quantizes the scattering coefficients into a given number of clusters. We then use these clusters to define a set of distances that can subsequently be used for similarity retrieval. Vincent : The verb use is repeated in the previous sentence. Testing 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 subsequently reported in Section 7.

2 Background

Computational bioacoustics usually 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 belonging to this paradigm attempt to infer global properties of bioacoustic scenes, including biodiversity indices [**Bardeli2010**], migration patterns [**Obrist2010**], as well as marking time intervals of particular interest for detailed human inspection [**rosenstock2002landbird**].

In the closely related field of urban sound environment analysis, a popular approach is the bag of frames (BOF) introduced in this context by Aucouturier et al. [**aucouturier2007bag**], where a scene is modeled by high-level summary statistics computed from local features. Recently, the proposed implementation using Gaussian mixture models (GMMs) of mel-frequency cepstral coefficients (MFCCs) has been demonstrated to perform comparably to a direct averaging of the features for a variety of tasks [**lagrange:hal-01082501**]. This contrasts with the typical morphology of acoustic scenes, which we believe is a "skeleton of events on a bed of textures" [**nelken2013**] where a few discrete sound events are superimposed upon a stationary acoustic background. As those distinct and sparse events are often highly relevant for the task at hand, it is arguably more beneficial to focus on characterizing these distinct events through describing the large-scale temporal evolution of auditory scenes rather than on summary statistics of short-term features.

This statement has some support in auditory psychology as well as sound synthesis based on summary statistics [**mcdermott2013summary**]. For example, studies in the cognitive psychology of urban sound environments have shown that global sound level (perceived or measured) is not sufficient to fully characterize an acoustic scene [**guyot2005urban**, **kang2006urban**]. Rather, it seems that cognitive processes such as sound environment quality perception [**dubois2006cognitive**] or loudness judgment [**kuwano˙memory˙2003**] rely upon higher-level cognitive attributes such as the identities of sound sources which compose 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, the presence of birds is very likely to be heard in parks in urban areas and so are 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 [**ricciardi2015sound**, **lavandier2006contribution**]. 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 approaches which treat each observation the same and do not incorporate temporal structure between features. A method that takes this knowledge into account could therefore have potential for great impact in the modeling of acoustic scenes, provided that the representation of the representation of those distinct events is rich enough.

3 Wavelet scattering

Scattering transforms are time-shift invariant representations of audio signals computed by applying auditory and modulation wavelet filter banks alternated with complex modulus nonlinearities. This section first explains the importance of local invariance to time-shifting and stability to time-warping in the representation of acoustic scenes, then describes how the scattering transform is designed to satisfy these properties while retaining high discriminative power.

3.1 Invariance and stability in audio signals

The notion of invariance to time-shifting plays an essential role in acoustic scene similarity retrieval as well as acoustic scene classification. Indeed, recordings may be shifted locally without affecting their similarity to other recordings. To discard this superfluous source of variability, signals are first mapped into a time-shifting invariant feature space. These features are then used to calculate similarities or to train a classifier. Since invariance is ensured by the features, it does not have to be learned during this last model construction step.

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 it is important for representation to be robust 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 band-pass filter of center 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 $(J_1Q_1 - 1)$, 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 center 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 [**Fastl2007**].

The wavelet transform of an audio signal $\boldsymbol{x}(t)$ is obtained by convolution with all wavelet filters. Applying 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 center 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 $\phi(t)$ with cut-off frequency $1/T$ (and approximate duration T), to get

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

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 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 center

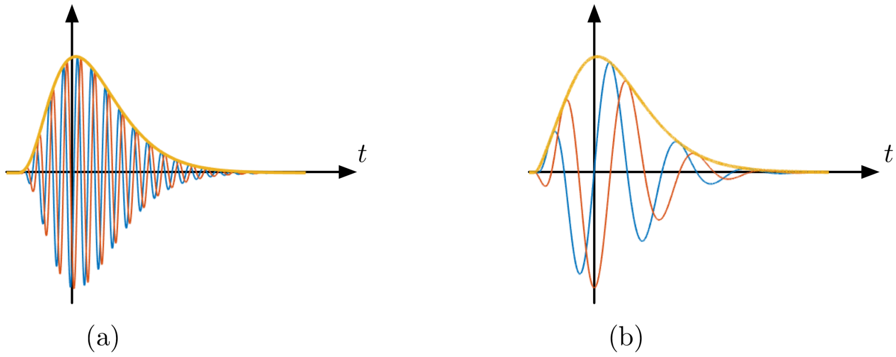

Fig. 1 Gammatone wavelets $\psi(t)$ in the time domain with quality factors (a) $Q = 4$ and (b) $Q = 1$. Oscillations (red, blue) are the real and imaginary parts. The envelope (yellow) is the complex modulus.

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 $\phi(t)$ 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

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

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 is a shorthand for 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 [**irene**].

3.4 Gammatone wavelets

Wavelets $\boldsymbol{\psi_{\gamma_1}}(t)$ and $\boldsymbol{\psi_{\gamma_2}}(t)$ are designed as fourth-order Gammatone wavelets with one vanishing moment [**Venkitaraman2014**], 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

[**Fastl2007**]. 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 [**Smith2006**]. 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 [**Hsu2003**]. 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.

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 linearizes 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

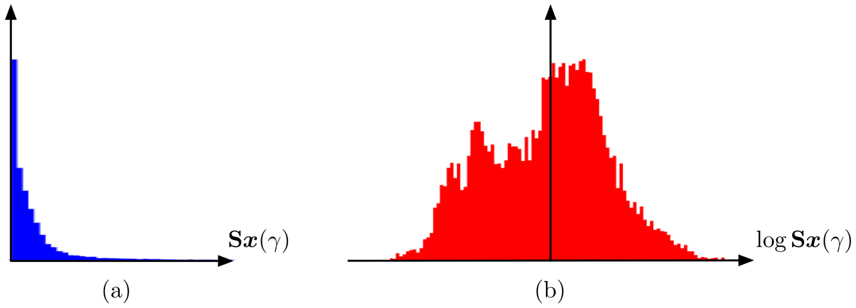

Fig. 2 Histogram of values taken by the first-order scattering coefficient $\mathbf{S}\boldsymbol{x}(\gamma)$, corresponding to a center acoustic frequency of 302 Hz, (a) before and (b) after logarithmic compression.

and unit variance to avoid mismatch in numeric ranges [**Hsu2003**]. 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 [**Anden2014**], this is not directly feasible in the unsupervised case. As the detection of events is still an open problem [**7100934**], 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. The partitioning is done by minimizing squared error between the empirical mean, or centroid, of each cluster and the vectors belonging to it. This is known as k-means clustering [**lloyd**] and has found widespread use for a variety of applications. Each scene s_u is then described by a set of clusters C_u. One should note that this quantization approach differs from unsupervised learning schemes such as the ones studied in [**bisot2016acoustic**], 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 [**selfTuneManor2004**]:

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

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, that is
$$\max_{m,n} K_{mn}^{uv}, \tag{7}$$
- 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, that is
$$\frac{1}{M^2}\sum_{m,n} K_{mn}^{uv} \tag{8}$$
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 [**pele2008linear**]. 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 with the "ground distance" between the histogram bins.

6 Experiments

To evaluate the representations introduced in the previous section, we apply it to the acoustic scene similarity retrieval task on the DCASE 2013 dataset.

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 [**7100934**] dataset. This dataset consists of two parts, namely 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 scence 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 DCASE 2013 dataset enjoys significant intra-class diversity while remaining of manageable size, making it suitable for extensive evaluation of algorithmic design choices [**lagrange:hal-01082501**].

6.2 Feature design

Experiments are carried out using scattering coefficients as well as baseline mel-frequency cepstral coefficients (MFCCs) as features. For the scattering transform, each 30-second scene is described by 128 vectors computed with half-overlapping windows $\boldsymbol{\phi}(t)$ of duration $T = 372\,\mathrm{ms}$, for a total of 24 s. In this case, 3 seconds are discarded at the beginning and end of the scene to avoid boundary artifacts. Experiments are conducted with and without logarithmic compression (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. The coefficients are averaged using 250 ms long non-overlapping windows so that each window represents structures of similar scale to the scattering coefficients.

6.3 Algorithm

The evaluation is performed on the private part of the DCASE 2013 dataset. The metric used is the precision at rank k ($p@k$), which is computed 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. These closest

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

neighbors are determined by a particular distance, which is one of *RbQ-c*, *RbQ-a*, or *RbQ-w*. Vincent : Avoid passive voice.

The metric $p@k$ is computed for $k = \{1, \ldots, 9\}$, since each class only has 10 items. Note that a $p@1$ is equivalent to the classification accuracy obtained by the classifier which chooses the label of the closest neighbor for a given item. The *RbQ* approaches 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 parameters of the BOF approach are the one used in [**aucouturier2007bag**], *i.e.* 10 Gaussians per scene estimated using the Expectation Maximization (EM) algorithm. Gaussian models are compared using Monte-Carlo sampling approaximation. In order to ensure convergence of the EM algorithm when considering the scattering features as input, the dimension of the feature vector is reduced from 1367 to 30 by projecting the features over the 30 principal components computed over the whole dataset. Vincent : Too much passive voice, too many instances of "use", "used".

The scaling parameter q of the RBF kernels (see Eq. 6) is set to 10% of the number of data points to cluster.The clustering for the *RbQ* approaches is done using plain k-means with random initialization and 200 replicates. For each method, the numbers of clusters is set to 8.

7 Results

This section presents evaluation results for the acoustic scene similarity retrieval task. The $p@k$ for different settings are shown in Figure 5, illustrating the effect of the different similarity metrics. This allows us to also evaluate the recall capabilities of the system under evaluation.

Baselines

As can be seen on Figure 3, the BOF approach behaves similarly to the early approach when considering MFCCs as features. Its performance decreases compared to the early approach when considering the scattering features. Extensive analysis of those results is out of the scope of this paper, but we can assume that this decrease is to to the dimensionality reduction operated on the scattering features before GMM training in order to ensure the convergence of the EM algorithm.

The early approach being simpler in terms of implementation and runtime complexity, this method is retained as baseline for the remaining.

Logarithmic compression

As can be seen on Figure 4, the logarithmic compression is beneficial for precision performance irrespective of the similarity considered. The gain is more pronounced when considering alternative similarities.

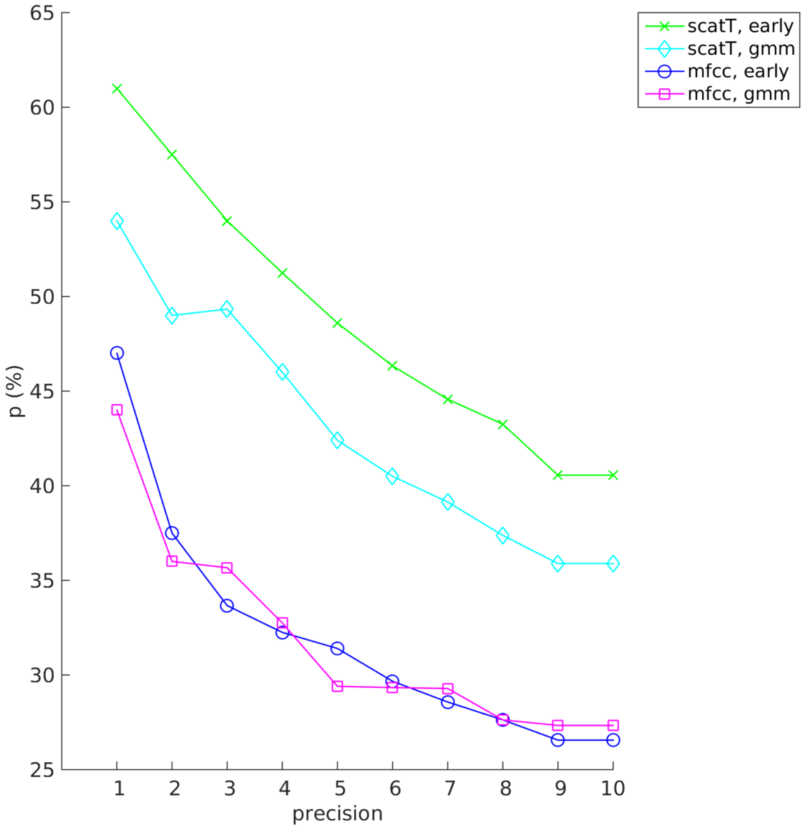

Fig. 3 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k ($p@k$) obtained for scattering with or without logarithmic compression, as a function of the rank k. Vincent : The x-axis label is wrong. precision should be replaced by "rank k". The y-axis should be "precision ($p@k$)". It would be good to have the y-axis range from 0% to 100%. The lines should be thicker and follow a color palette of complementary colors (blue vs. orange). The legends should not be abbreviated: scatT -¿ scattering, mfcc -¿ MFCC.

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.

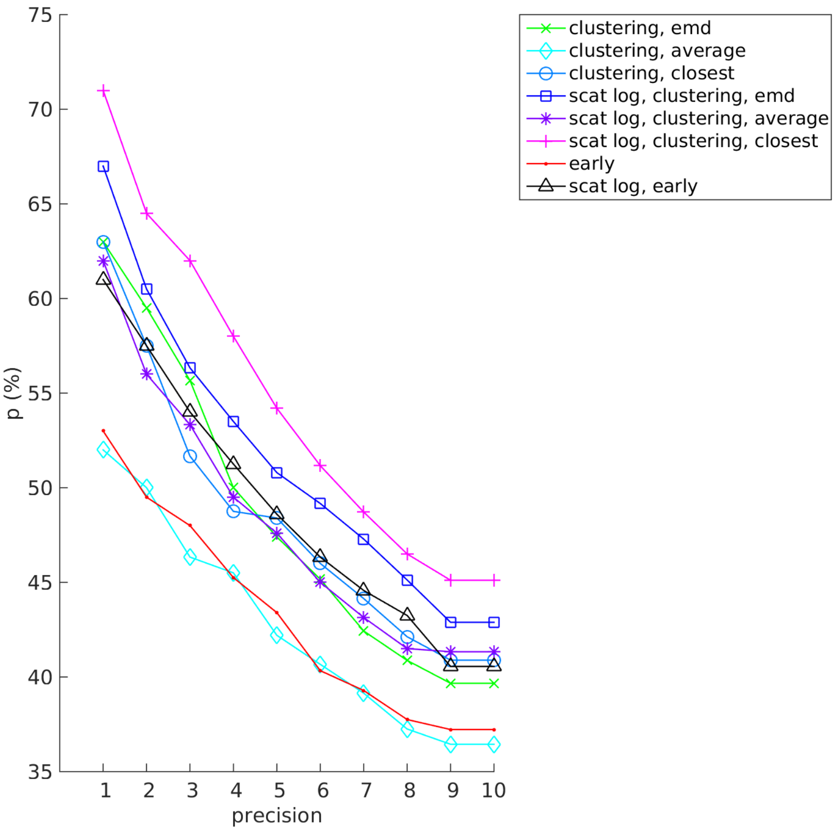

Fig. 4 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k ($p@k$) obtained for scattering with or without logarithmic compression, as a function of the rank k.

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 noticing that *RbQ-a* performs worse or comparably to *early*, showing that the discriminant information is destroyed by averaging the contributions from all centroids. This result is in line with the findings of [**lagrange:hal-01082501**]. 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 characterize the classes compared to *RbQ-w*. This last observation suggests that weighting a centroid according to the number of

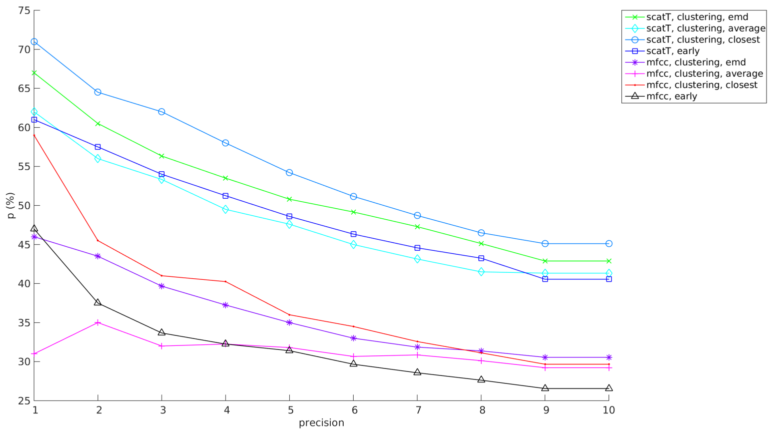

Fig. 5 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k ($p@k$) obtained for MFCCs and scattering with logarithmic compression, as a function of the rank k.

frames it contains may prove to be a limited solution. 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. Mathieu : This is a bit of a stretch, since it may just be that we used the wrong distance to take the weighting into account. Perhaps the non-normalized EMD is not appropriate here?

Joakim : needs to be discussed

Considering $p@5$ as our metric (as in [**aucouturier2007bag**] and [**lagrange:hal-01082501**]), the use of the log-scattering transform instead of MFCCs increases the performance from 0.31 to 0.49 and the use of the relevance-based quantization approach using the closest similarity (RbQ-c) further improves the performance to 0.54 for a global increase of 0.23. Vincent : Too many instances of “use”, “use”, “using”. It is worth reminding what the baselines are.

8 Conclusion

This paper presents a new approach for modeling acoustic scenes which utilizes scattering transforms at small scales and a cluster-based representation at large scales. Compared to traditional models based on BOF and summary statistics, 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. In order to extend the selectivity potential of late integration using SVMs to the unsupervised case, we develop the 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 [**wimmer2013sampling**]. Vincent : The verb utilize should be avoided.