Dear Author,

Here are the final proofs of your article. Please check the proofs carefully.

All communications with regard to the proof should be sent to SpringerOpen_Production@spi-global.com.

Please note that at this stage you should only be checking for errors introduced during the production process. Please pay particular attention to the following when checking the proof:

- Author names. Check that each author name is spelled correctly, and that names appear in the correct order of first name followed by family name. This will ensure that the names will be indexed correctly (for example if the author's name is 'Jane Patel', she will be cited as 'Patel, J.').
- Affiliations. Check that all authors are cited with the correct affiliations, that the author who will receive correspondence has been identified with an asterisk (*), and that all equal contributors have been identified with a dagger sign (†).
- Ensure that the main text is complete.
- Check that figures, tables and their legends are included and in the correct order.
- Look to see that queries that were raised during copy-editing or typesetting have been resolved.
- Confirm that all web links are correct and working.
- Ensure that special characters and equations are displaying correctly.
- Check that additional or supplementary files can be opened and are correct.

Changes in scientific content cannot be made at this stage unless the request has already been approved. This includes changes to title or authorship, new results, or corrected values.

How to return your corrections

Returning your corrections via online submission:
- Please provide details of your corrections in the online correction form. Always indicate the line number to which the correction refers.

Returning your corrections via email:
- Annotate the proof PDF with your corrections.
- Send it as an email attachment to: SpringerOpen_Production@spi-global.com.
- Remember to include the journal title, manuscript number, and your name when sending your response via email.

After you have submitted your corrections, you will receive email notification from our production team that your article has been published in the final version. All changes at this stage are final. We will not be able to make any further changes after publication.

Kind regards,

SpringerOpen Production Team

Relevance-based quantization of scattering features for unsupervised mining of environmental audio

Vincent Lostanlen, Grégoire Lafay, Joakim Andén and Mathieu Lagrange[*]

Abstract

The emerging field of computational acoustic monitoring 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 further, 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 interest 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 activity and climate change on animal biodiversity [34], researchers have recently undertaken a massive deployment of acoustic sensors throughout the world [27, 33, 37]. In addition, recent work has explored acoustic monitoring for characterization of human pleasantness in urban areas [11, 29], as well as the prediction of annoyance due to traffic [10]. Since they bear a strong societal impact and raise many scientific challenges, we believe that these applications are of considerable interest for signal processing community.

An important problem is that manually analyzing the recorded data to identify the quantities of interest is very costly. Some sort of pre-screening is therefore required to reduce the need for human expert listening and annotation. To this aim, the most straightforward approach is to specify a closed set of sound classes, such as sound classes expected to appear near the acoustic sensors. Computational models are then trained for these classes which are used to automatically annotate recordings [40]. 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 prediction on unseen data—i.e., sensors outside of the training set—is prone to errors. Secondly, and more importantly, it is based on prior knowledge and thus cannot be considered for exploratory analysis, in which quantities of interest have yet to be defined.

To identify which parts need human inspection, one needs tools that are able to detect both recurring patterns

*Correspondence: mathieu.lagrange@cnrs.fr
LS2N, Nantes, France

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. To construct such an algorithm, we are required to represent an audio recording in a way that captures its distinctive qualities. A widespread choice of representation is the bag of frames [3], which describes an auditory scene recording using summary statistics of short-time features. Unfortunately, the bag-of-frames approach 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 sparsely distributed over time. Furthermore, experiments in cognitive psychology [9] and cognitive neuroscience [25] suggest that human acoustic perception is highly sensitive to such isolated sound events. We believe that the failure to model such distinct events is one of the reasons why the bag-of-frames representation is insufficient [15].

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 [7], MFCCs have recently found wider use in music information retrieval [19] and environmental audio processing [3]. A richer representation, the scattering transform, has enjoyed significant success in various audio [1] and biomedical [6] signal classification tasks. Its structure is that of a convolutional neural network [2, 4, 17, 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 that it does not require 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 a cluster model. This unsupervised model quantizes the scattering coefficients into a given number of clusters. These clusters are then used 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

Characterization of the similarity between audio recordings be they at the scale of the minute, the hour, the day or larger is of interest for many application areas involving acoustic monitoring such as urban sound environment analysis and ecoacoustics. In this context, a classical approach is the bag of frames (BoF), first applied to the problem by Aucouturier et al. [3]. It models an auditory scene using high-level summary statistics computed from local features, typically implemented by Gaussian mixture models (GMMs) of MFCCs.

It is worth mentionning that this task typically falls into an unsupervised paradigm where no prior knowledge is used to model a given scene. For each scene s, a model M_s is computed. The similarity among the scene s_1 and s_2 is computed as the similarity (see Section 6 for further details). The BoF approach is also widely used in a supervised fashion for solving a classification task [32]. In this case, each class of scenes from a given typology, say {*park, boulevard, square*}, is modeled by a GMM trained on scenes taken from a training set. In order to predict the class of a given scene s, the likelihood each model given the scene are computed. The scene is then labeled park if the likelihood of the GMM trained on park scenes is higher than all other likelihoods.

While BoF has largely been superseded by more sophisticated methods for the task of acoustic scene classification [32], it remains the best-performing model for acoustic scene similarity retrieval, though this representation was recently shown to perform comparably to direct averaging of the features for a variety of datasets [15]. 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 [23]. 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 [11, 13]. Instead, cognitive processes

such as sound environment quality perception [9] or loudness judgment [14] 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 [16, 29]. 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 $x(t)$, we would like its translation $x_c(t) = x(t - c)$ 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 $x_\tau(t) = x(t - \tau(t))$ for some function τ, which performs a time-warping of $x(t)$ to obtain $x_\tau(t)$. 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 $x(t)$ and $x_\tau(t)$ are close together if $|\tau'(t)|$ is bounded by a small constant [22].

3.2 Wavelet scalogram

Our convention for the Fourier transform of a continuous-time signal $x(t)$ is $\hat{x}(\omega) = \int_{-\infty}^{+\infty} x(t) \exp(-i2\pi \omega t) \, dt$. Let $\psi(t)$ be 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 $\psi(t)$ according to a geometric sequence of scales $2^{\gamma_1/Q_1}$, obtaining

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

Q5

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 $\psi_{\gamma_1}(t)$ 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 ψ. 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 [41].

The wavelet transform of an audio signal $x(t)$ is obtained by convolution with all wavelet filters. Applying a pointwise complex modulus, the transform yields the wavelet scalogram

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

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 ψ_{γ_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

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

which is known as the set of first-order scattering coefficients. They capture the average spectral envelope of $x(t)$ 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 $x_1(t, \gamma_1)$ into the time-shift invariant representation $\mathbf{S_1}x(t, \gamma_1)$, they can be recovered from $x_1(t, \gamma_1)$ by a second wavelet transform and another complex modulus.

We define second-order wavelets $\psi_{\gamma_2}(t)$ 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

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

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 $x_2(t, \gamma_1, \gamma_2)$ to guarantee local invariance to time-shifting, which yields the second-order scattering coefficients

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

The scattering transform $\mathbf{S}x(t, \gamma)$ consists of the concatenation of first-order coefficients $\mathbf{S_1}x(t, \gamma_1)$ and second-order coefficients $\mathbf{S_2}x(t, \gamma_1, \gamma_2)$ into a feature matrix $\mathbf{S}x(t, \gamma)$, 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 [36].

3.4 Gammatone wavelets

Wavelets $\psi_{\gamma_1}(t)$ and $\psi_{\gamma_2}(t)$ are designed as fourth-order Gammatone wavelets with one vanishing moment [35] and are shown in Fig. 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 [41]. 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 masking tone has to be louder if it is above the stimulus in frequency rather than below. It should also be noted that Gammatone
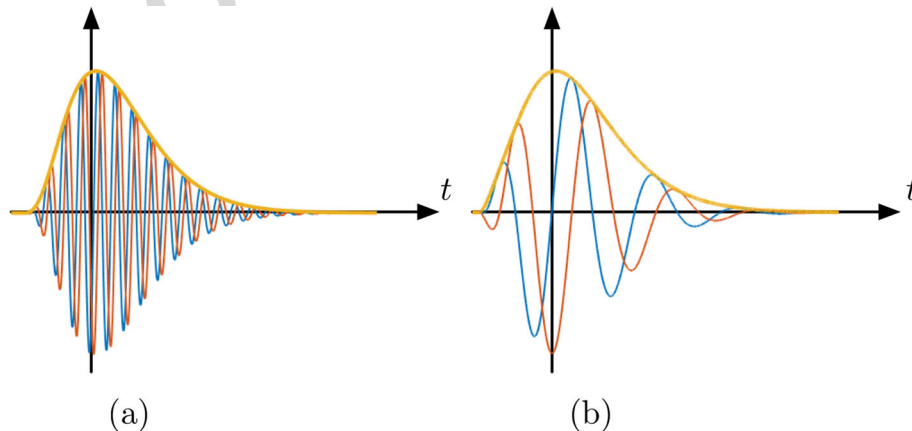

Fig. 1 Gammatone wavelets $\psi(t)$ in the time domain with quality factors **a** $Q = 4$ and **b** $Q = 1$. Oscillations (red, blue) are the real and imaginary parts. The envelope (yellow) is the complex modulus

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 [30, 31]. In addition, Gammatone wavelets have proven to outperform Morlet wavelets on a benchmark of supervised musical instrument classification from scattering coefficients [20]. 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 support vector machines, tend to work best when all features follow a standard normal distribution across all training instances [12]. 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 linearize this dependency, facilitating the construction of powerful invariants at the classifier stage.

4.2 Standardization

Let $\mathbf{S}x(\gamma, n)$ 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 [12]. To standardize $\mathbf{S}x(\gamma, n)$, we subtract the sample mean vector $\mu[\mathbf{S}x(\gamma)]$ from $\mathbf{S}x(\gamma, n)$ and divide the result by the sample standard deviation vector $\sigma[\mathbf{S}x](\gamma)$. The vectors $\mu[\mathbf{S}x(\gamma)]$ and $\sigma[\mathbf{S}x](\gamma)$ 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 [1], this is not directly feasible in the unsupervised case. As the detection of events is still an open problem [32], 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 [18]. 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 [5], where the scene
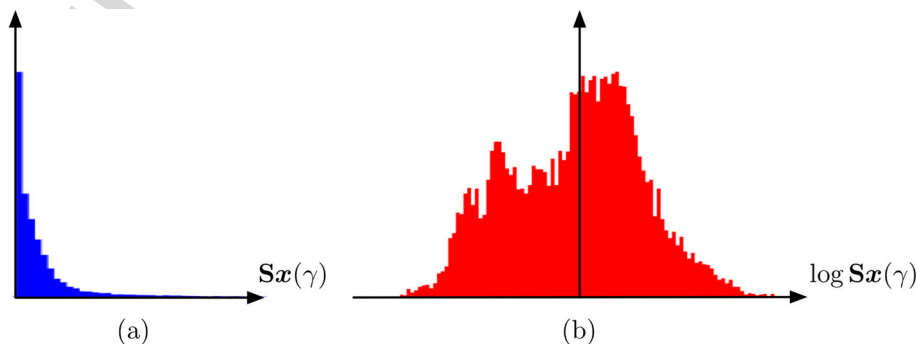

Fig. 2 Histogram of values taken by the first-order scattering coefficient $\mathbf{S}x(\gamma)$, corresponding to a central acoustic frequency of 302 Hz, **a** before and **b** after logarithmic compression

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]:

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

Here, $\mu_{m,q}^u$ and $\mu_{n,q}^v$ are the qth nearest neighbors to the centroids μ_m^u and μ_n^v, respectively, and the double bars $\|\cdot\|$ denote 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

$$\max_{m,n} K_{mn}^{uv}, \qquad (7)$$

- 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

$$\frac{1}{M^2} \sum_{m,n} K_{mn}^{uv} \qquad (8)$$

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

$$p_u = \left\{\left(\mu_1^u, w_1^u\right), \left(\mu_2^u, w_2^u\right), \ldots, \left(\mu_M^u, w_M^u\right)\right\}, \qquad (9)$$

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 [28]. 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 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 publicly available DCASE 2013 dataset [32]. Although the dataset was constructed for the task of acoustic scene classification, where the goal is to correctly assign the class of a given recording, we can use the same recordings and class labels for the task of acoustic scene similarity retrieval. The 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 s in duration. The dataset is evenly divided into 10 acoustic scene classes: *bus*, *busy street*, *office*, *open air market*, *park*, *quiet street*, *restaurant*, *supermarket*, *tube*, and *tube station*. 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. In order to avoid any correlation between recording conditions and label distribution, all recordings were carried out 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 [15]. As an illustration, Fig. 3 represents the wavelet scalogram of one recording within the DCASE 2013, labeled as *park*. **F3**
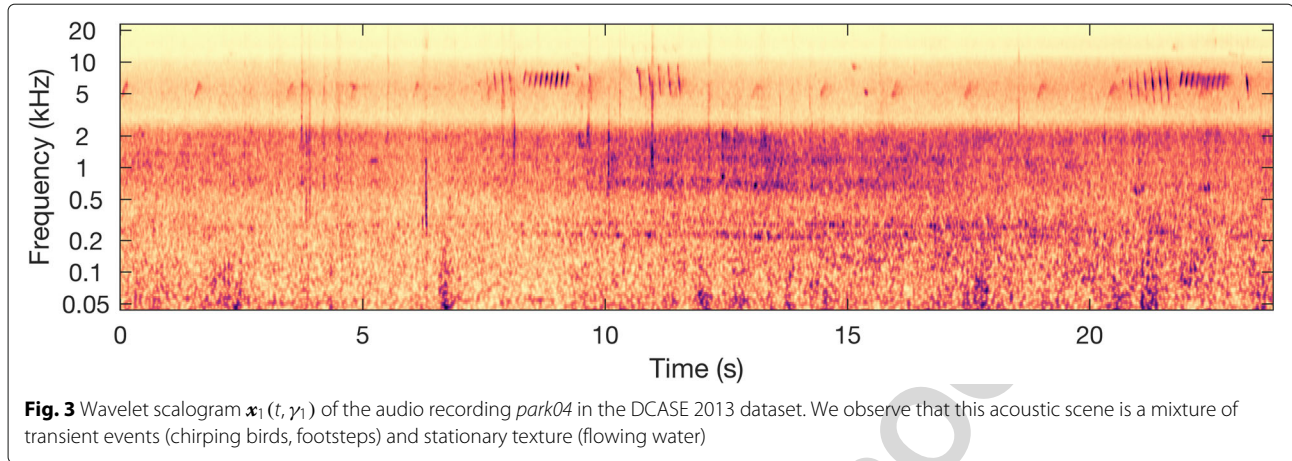
6.2 Feature design

We perform our experiments using both scattering coefficients and MFCCs. For the scattering transform, each 30-s scene is described by 128 vectors of dimension 1367 computed with half-overlapping windows $\boldsymbol{\phi}(t)$ of duration $T = 372$ ms, for a total of 24 s. Here, we discard 3 s 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 Evaluation and algorithm

The evaluation is performed on the private part of the DCASE 2013 dataset. As a metric, we use the precision

Fig. 3 Wavelet scalogram $x_1(t, \gamma_1)$ of the audio recording *park04* in the DCASE 2013 dataset. We observe that this acoustic scene is a mixture of transient events (chirping birds, footsteps) and stationary texture (flowing water)

at rank k (*p@k*). This number 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. We determine these neighbors using one of the proposed similarity measures *RbQ-c*, *RbQ-a*, or *RbQ-w*. We compute *p@k* for $k = \{1, \dots, 9\}$, 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. [3], GMMs are estimated for each scene using the expectation-maximization (EM) algorithm [8, 24]. 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 [3] 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. As the number of Gaussians for the BoF approach, the numbers of clusters M controls the level of abstraction. For each method, unless otherwise stated, the parameter M is set to 8. It thus allows 8 different types of observations to be modeled, which seems reasonable given the duration of the scene (30 s). Note that this is the only free parameter in the proposed method. However, except for *RbQ-a*, the results are not very sensitive to the choice of M, as long

as it is large enough to characterize the seen. A numerical demonstration is provided at the end of the next section.

7 Results

Results for the acoustic scene similarity retrieval task demonstrate that logarithmically compressed scattering features outperform MFCCs. Combining these with the *RbQ* cluster model, improvements are obtained over traditional BoF and summary statistic measures.

7.0.1 Baselines

As seen in Fig. 4, scattering features significantly outperform the baseline MFCCs for both BoF and early integration schemes. This is expected, as the scattering transform extends the MFCCs by including supplementary amplitude modulation information [1]. We also note that applying PCA reduction to the scattering transform has little effect for the early integration scheme. In the context of summary statistics, 30 dimensions are sufficient to discriminate between auditory scenes.
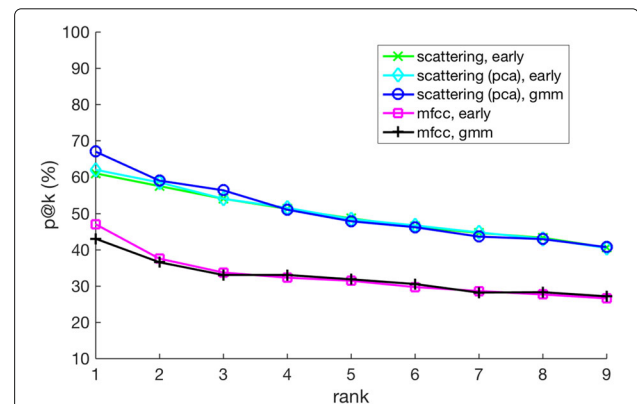

Fig. 4 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k (*p@k*) obtained for several baseline approaches, as a function of the rank k

Comparing the BoF to early integration, both approaches perform similarly for MFCCs and PCA-reduced scattering features alike. This is in line with previous results on BoF [15], where it is found to perform similarly as similarity retrieval on features averaged over the entire recording.

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

7.0.2 Logarithmic compression

Figure 5 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.

7.0.3 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 (Fig. 6). 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.

7.0.4 Relevance-based quantization vs. early integration

For the scattering transform, both *RbQ-c* and *RbQ-w* outperform *early*, thus confirming the benefits of using a 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 [15]. To take advantage of such a representation, we need to select certain representative centroids when comparing quantized objects. The same behavior is observed for MFCCs,
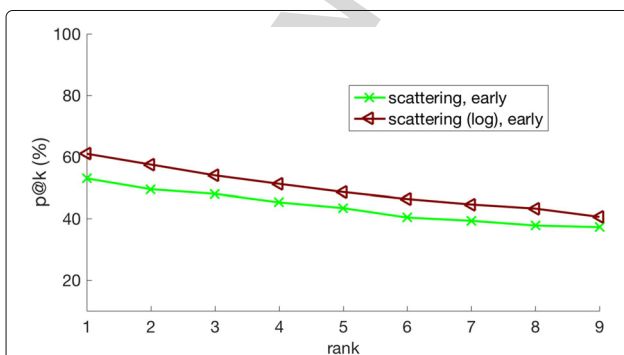

Fig. 6 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k (*p@k*) obtained for MFCCs and scattering with logarithmic compression

with *RbQ-c* and *RbQ-w* outperforming *early*, which is equivalent to the state-of-the-art BoF model, as seen previously.

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. [3] and [15]), 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.

7.0.5 Sensitivity to number of clusters M

We now study the sensitivity of the precision at rank 5 (*p@5*) with respect to the number of clusters M. The results are shown in Fig. 7.

For a small number of clusters ($M = 1$ or $M = 2$), all methods perform worse, since not enough discriminative sound objects are extracted from the recording. Please note that setting $M = 1$ is equivalent to the *early* approach as this corresponds a summary statistics model. For $M = 4$, most methods perform well, since this allows for better characterization of various signal structures in

Fig. 5 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
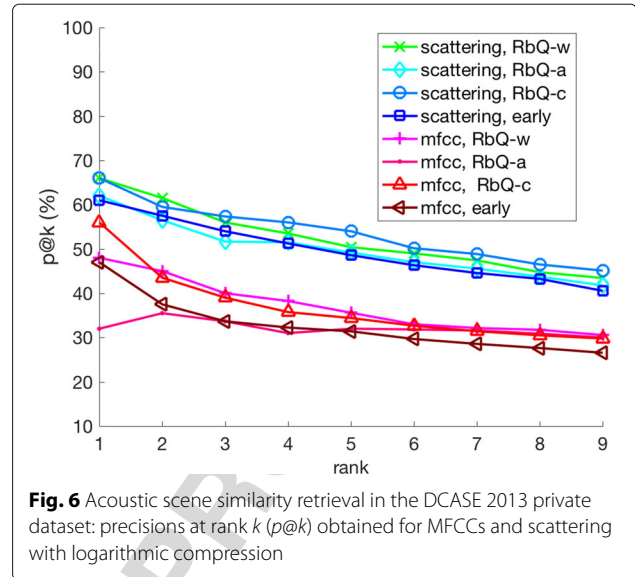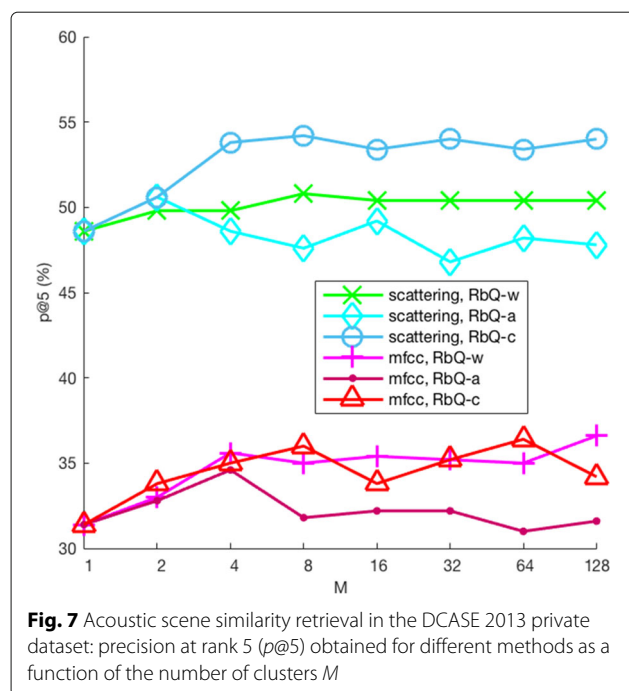
Fig. 7 Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precision at rank 5 (*p*@5) obtained for different methods as a function of the number of clusters *M*

the scenes. As the number of clusters increases, the *RbQ-a* method performs worse for both scattering features and MFCCs since any distinct objects are averaged out by clusters representing the background. The *RbQ-w* and *RbQ-c* methods do better in this regard, as they are better able to emphasize the clusters that discriminate well between scenes.

Using *RbQ-c*, therefore, we are not very sensitive to the choice of *M*, as long as it is large enough to allow for separation of the discriminative sound objects from the background. This motivates our choice of $M = 8$ for the previous experiments in this section.

8 Conclusions

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 ecoacoustics [34, 38].

Endnote

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

Q9

Additional file

Q10
Q11

Additional file 1: Supplementary material. (PDF 63 KB)

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:

- Companion website: https://github.com/mathieulagrange/paperRelevanceBasedSimilarity
- Programming language: Matlab
- License: GNU GPL
- Any restrictions to use by non-academics: license needed

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.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. J. Andén, S. Mallat, Deep scattering spectrum. IEEE Trans Sig Process. **62**(16), 4114–4128 (2014)
2. R. Arandjelović, A. Zisserman, Look, listen and learn (2017)
3. J. J. Aucouturier, B. Defreville, F. Pachet, The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. J. Acoust. Soc. Am. **122**(2), 881–891 (2007) — Q12
4. Y. Aytar, C. Vondrick, A. Torralba, in *Advances in Neural Information Processing Systems*. Soundnet: Learning sound representations from unlabeled video, (2016), pp. 892–900 — Q13
5. V. Bisot, R. Serizel, S. Essid, G. Richard, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Acoustic scene classification with matrix factorization for unsupervised feature learning (IEEE, 2016), pp. 6445–6449 — Q14
6. V. Chudáček, J. Anden, S. Mallat, P. Abry, M. Doret, Scattering transform for intrapartum fetal heart rate variability fractal analysis: A case study. IEEE Trans. Biomed. Eng (2013) — Q15
7. S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Sig Process. **28**(4), 357–366 (1980)
8. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. B Stat. Methol. **39**(1), 1–38 (1977). http://www.jstor.org/stable/2984875
9. D. Dubois, C. Guastavino, M. Raimbault, A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories. Acta Acustica U with Acustica. **92**(6), 865–874 (2006)
10. J. R. Gloaguen, A. Can, M. Lagrange, J. F. Petiot, in *Workshop on Detection and Classification of Acoustic Scenes and Events*. Estimating traffic noise levels using acoustic monitoring: a preliminary study, (2016)

11. F. Guyot, S. Nathanail, F. Montignies, B. Masson, in *Proceedings Forum Acusticum*. Urban sound environment quality through a physical and perceptive classification of sound sources: A cross-cultural study, (Budapest, Hungary, 2005)

12. 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)

13. J. Kang, *Urban sound environment*. (CRC Press, 2006)

14. S. Kuwano, S. Namba, T. Kato, J. Hellbrck, Memory of the loudness of sounds in relation to overall impression. Acoust Sci Technics. **4**(24) (2003)

15. M. Lagrange, G. Lafay, B. Defreville, J. J. Aucouturier, The bag-of-frames approach: A not so sufficient model for urban soundscapes. JASA Express Lett. **138**(5), 487–492 (2015)

16. C. Lavandier, B. Defréville, The contribution of sound source characteristics in the assessment of urban soundscapes. Acta Acustica U with Acustica. **92**(6), 912–921 (2006)

17. H. Lee, P. Pham, Y. Largman, A. Ng, in *Proc. NIPS*. Unsupervised feature learning for audio classification using convolutional deep belief networks, (2009)

18. S. Lloyd, Least squares quantization in PCM. IEEE Trans. Inf. Theory. **28**(2), 129–137 (1982). https://doi.org/10.1109/TIT.1982.1056489

19. B. Logan, in *Proceedings of the International Symposium on Music Information Retrieval*. Mel frequency cepstral coefficients for music modeling, (2000)

20. V. Lostanlen, *Convolutional operators in the time-frequency domain. Ph.D. thesis*. (École normale supérieure, 2017)

21. V. Lostanlen, C. E. Cella, in *Proceedings of the International Society for Music Information Retrieval Conference. ISMIR*. Deep convolutional networks in the pitch spiral for music instrument classification, (2016)

22. S. Mallat, Group Invariant Scattering. Commun. Pur. Appl. Math. **65**(10), 1331–1398 (2012). https://doi.org/10.1002/cpa.21413

23. J. H. McDermott, M. Schemitsch, E. P. Simoncelli, Summary statistics in auditory perception. Nat. Neurosci. **16**(4), 493–498 (2013)

24. T. K. Moon, The expectation-maximization algorithm. IEEE Signal Proc. Mag. **13**(6), 47–60 (1996)

25. I. Nelken, Processing of complex stimuli and natural scenes in the auditory cortex. Curr. Opin. Neurobiol. **14**(4), 474–480 (2004)

26. I. Nelken, A. de Cheveigné, An ear for statistics. Nat. Neurosci. **16**(4), 381–382 (2013)

27. S. R. Ness, H. Symonds, P. Spong, G. Tzanetakis, The Orchive: Data mining a massive bioacoustic archive. Int. Work. Mach. Learn. Bioacoustics (2013)

28. O. Pele, M. Werman, in *European Conference on Computer Vision*. A linear time histogram metric for improved SIFT matching (Springer, 2008), pp. 495–508

29. P. Ricciardi, P. Delaitre, C. Lavandier, F. Torchia, P. Aumond, Sound quality indicators for urban places in Paris cross-validated by Milan data. J. Acoust. Soc. Am. **138**(4), 2337–2348 (2015)

30. T. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, O. Vinyals, in *Proceedings of INTERSPEECH*. Learning the speech front-end with raw waveform cldnns, (2015)

31. E. C. Smith, M. S. Lewicki, Efficient auditory coding. Nature. **439**(7079), 978–982 (2006)

32. D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, M. D. Plumbley, Detection and classification of acoustic scenes and events. IEEE Trans. Multimed. **17**(10), 1733–1746 (2015)

33. D. Stowell, M. D. Plumbley, Large-scale analysis of frequency modulation in birdsong databases. Methods Ecol. Evol. **11** (2013)

34. J. Sueur, A. Farina, Ecoacoustics: the ecological investigation and interpretation of environmental sound. Biosemiotics. **8**(3), 493–502 (2015)

35. A. Venkitaraman, A. Adiga, C. S. Seelamantula, Auditory-motivated Gammatone wavelet transform. Sig. Process. **94**, 608–619 (2014)

36. I. Waldspurger, in *Proc. SampTA*. Exponential decay of scattering coefficients, (2017), pp. 143–146. https://doi.org/10.1109/SAMPTA.2017.8024473

37. P. S. Warren, M. Katti, M. Ermann, A. Brazel, Urban bioacoustics: It's not just noise. Anim. Behav. **71**(3), 491–502 (2006)

38. J. Wimmer, M. Towsey, P. Roe, I. Williamson, Sampling environmental acoustic recordings to determine bird species richness. Ecol. Appl. **23**(6), 1419–1428 (2013)

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

40. L. Zhang, M. Towsey, J. Zhang, P. Roe, Classifying and ranking audio clips to support bird species richness surveys. Ecol. Inform. **34**, 108–116 (2016). https://doi.org/10.1016/j.ecoinf.2016.05.005. https://eprints.qut.edu.au/96243/

41. E. Zwicker, H. Fastl, *Psychoacoustics: Facts and models, vol. 22*. (Springer Science & Business Media, 2013)

Author Query Form

Journal: EURASIP Journal on Audio, Speech, and Music Processing

Article: Relevance-based quantization of scattering features for unsupervised mining of environmental audio

Dear Author,

During the copyediting of your paper, the following queries arose. Please respond to these by annotating your proofs with the necessary changes/additions.

- If you intend to annotate your proof electronically, please refer to the E-annotation guidelines.
- If you intend to annotate your proof by means of hard-copy mark-up, please refer to the proof mark-up symbols guidelines. If manually writing corrections on your proof and returning it by fax, do not write too close to the edge of the paper. Please remember that illegible mark-ups may delay publication.

Whether you opt for hard-copy or electronic annotation of your proofs, we recommend that you provide additional clarification of answers to queries by entering your answers on the query sheet, in addition to the text mark-up.

Query No.	Query	Remark
Q1	Author names: Please confirm if the author names are presented accurately, and in the correct sequence (given names/initials, family name). Author 1 Given Name: Vincent Last Name: Lostanlen Author 2 Given Name: Grégoire Last Name: Lafay Author 3 Given Name: Joakim Last Name: Andén Author 4 Given Name: Mathieu Last Name: Lagrange	
Q2	Please check if the provided city for affiliation 1 is correct. Otherwise, please amend.	
Q3	Please check if the edit to the sentence "For each scene s, a model M_s is computed." is correct.	
Q4	Please check if the edit to the sentence "While BoF has largely been superseded by more sophisticated methods for the task…" is correct.	
Q5	Equations: Please check all equations to ensure that they are captured/presented correctly. Otherwise, kindly amend if necessary.	
Q6	Figure: Please check and confirm if all captions and citations of figures have been captured correctly.	
Q7	Please check if the edit to the sentence " In the early approach being simpler in terms of implementation and runtime complexity…" is correct.	
Q8	Missing citation for Figure 6 was inserted in the sentence "Overall, log-compressed scattering coefficients…." Please check if appropriate. Otherwise, please provide a citation for Figure 6. Note that the order of main citations of figures in the text must be sequential.	

Query No.	Query	Remark
Q9	Please verify if the following URLs are working: https://github.com/mathieulagrange/paperRelevanceBasedSimilarity; http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/description.html; https://github.com/mathieulagrange/paperRelevanceBasedSimilarity; http://www.jstor.org/stable/2984875; https://doi.org/10.1109/TIT.1982.1056489; https://doi.org/10.1002/cpa.21413; https://doi.org/10.1109/SAMPTA.2017.8024473; https://doi.org/10.1016/j.ecoinf.2016.05.005; https://eprints.qut.edu.au/96243/; https://github.com/mathieulagrange/paperRelevanceBasedSimilarity; http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/description.html; http://www.jstor.org/stable/2984875; https://eprints.qut.edu.au/96243/. Otherwise, please provide alternatives.	
Q10	Additional file: Supplementary material section found in the manuscript was captured as separated Additional file 1 in pdf format. Please check and advise if the capturing of additional file references are correct.	
Q11	Additional file: Additional file 1 was received; however, no citation was provided in the manuscript. Please provide the location of where to insert the citation in the main body of the text. Otherwise, kindly advise us on how to proceed. Please note that additional files should be cited in ascending numerical order in the main body of the text.	
Q12	Please provide a complete bibliographic details for reference [2].	
Q13	References: Citation details for References [4, 12, 20, 22, 25, 36, 42] are incomplete. Please supply the Publisher name and publisher address of these references. Otherwise, kindly advise us on how to proceed.	
Q14	References: Citation details for References [5, 7, 16, 24, 33, 34, 47] are incomplete. Please supply the Publisher address of these references. Otherwise, kindly advise us on how to proceed.	
Q15	References: Citation details for References [6, 31] are incomplete. Please supply the Volume number and page range of these references. Otherwise, kindly advise us on how to proceed.	
Q16	References: Citation details for Reference [14] are incomplete. Please supply the Publisher name of this reference. Otherwise, kindly advise us on how to proceed.	
Q17	References: Citation details for References [17, 39] are incomplete. Please supply the Page range of these references. Otherwise, kindly advise us on how to proceed.	