Auditory Scene Similarity Retrieval and Classification with Relevance-based Quantization of Scattering Features

Vincent Lostanlen, Grégoire Lafay, Joakim Andén, and Mathieu Lagrange

***Abstract*—The emerging field of computational ecoacoustics aims at retrieving high-level information from the acoustic scenes recorded by some network of sensors. In this article, we address the problem of acoustic scene classification according to the taxonomy of the DCASE 2013 challenge, both in supervised and unsupervised settings. Traditional approaches compute a ”bag-of-frames”, i.e. 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. We validate the use of the scattering transform for supervised acoustic scene classification by reporting a better accuracy than MFCCs when using a support vector machine (SVM) classifier. Furthermore, we illustrate the power of the full cluster-based representation by applying it to an unsupervised acoustic scene similarity retrieval task, where we significantly outperform more traditional approaches based on summary statistics of MFCCs.**

***Index Terms*—acoustic scene classification, acoustic scene similarity retrieval, cluster models, wavelets, scattering, invariance.**

I. 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 the 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]. Because 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 works 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 [7, 8]. Despite identified needs, the field of signal-based ecoacoustics is still in its infancy. Consequently, few well-designed datasets are readily available for evaluation purposes.

An important problem in this field is the characterization of auditory scenes for the purposes of similarity estimation. One application of this is acoustic scene similarity retrieval, where the scenes judged “most similar” to a given query scene are extracted from some larger dataset. This requires the auditory scenes to be represented in some way that captures their distinctive qualities. Often, this has been achieved by bag-of-frames approaches [9], which attempts 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 scenes which are dynamic and contain episodic, distinct sound events that discriminate the scene from others. Indeed, experiments in cognitive psychology [10] and cognitive neurosciences [11] suggest that this matches more closely the way that humans perceive acoustic scenes.

To design better representations for the similarity retrieval task, we draw inspiration from a closely related, and more mature, field of investigation: the classification of acoustic scenes. This task, consisting of predicting scene labels from audio recordings, is narrower in its range of applications, but has the advantage of being rooted in the cognitive psychology of categorization [10, 12]. In addition, it has received significant attention from the data mining community for some time, resulting in the availability of well-designed datasets and competitive algorithms for solving it.

Adressing this task requires first to be able 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 [13], but which have found wider use in recent years in music information retrieval [14] and environmental audio processing [9]. A richer set of features are found in the scattering transform, a generic signal representation which has enjoyed significant success in various audio classification tasks [15].

The next question is the choice of a classifier. Here, a class of particularly successful algorithms train a classifier, such as a support vector machine (SVM) [16], to discriminate between short extracts of an auditory scene, and at the classification stage divide the signal into short extracts which are classified separately before pooling the classification results to reach a final decision. This last step is known as late integration and is crucial to the success of these methods, since it allows the classifier to concentrate on the short,

This work is supported by the ERC InvariantClass grant 320959.

distinct sound events that discriminate between scenes. The success of the late integration technique again leads credence to the idea that rare, distinct events in an auditory scene are sufficient to characterize it almost completely. Combining scattering transforms for low-level representation with late-integrated SVMs for classifying the scene at a higher level, we obtain a competitive auditory scene classification system, demonstrating the usefulness of both concepts for this task.

Returning to scene retrieval setting, we cannot use supervised models such as SVMs, but we can emulate their selection behavior by quantizing the auditory scene and defining a distance over this quantization. We therefore 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. Testing this approach on an 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 the field of acoustic scene modeling is given in Section II. We describe the scattering transform in Section III, discuss processing of the resulting features in Section IV and propose a cluster-based scene description in Section VI. In Section VII, several experiments are described for the acoustic scene classification task and the acoustic scene similarity retrieval task. Results are subsequently reported in Section VIII.

II. Background

Computational bioacoustics usually refers to the investigation of sound production, dispersion and reception in animals (including humans) by computational means. 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 [17], migration patterns [18], as well as marking time intervals of particular interest for detailed human inspection [19].

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. [9], 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 [20]. This contrasts with the typical morphology of acoustic scenes, which we believe is a "skeleton of events on a bed of textures" [21] where a few discrete sound events are superimposed upon a stationary acoustic background. As those distinct and scarce events are most of the time 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 the joint probability distribution of short-term features.

This statement has some support in auditory psychology as well as sound synthesis based on summary statistics [22]. 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 [4, 23]. Rather, it seems that cognitive processes such as sound environment quality perception [10] or loudness judgment [24] 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 [5, 25]. 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 equally when computing summary statistics. A method that takes this knowledge into account could therefore have potential for great impact in the modeling of acoustic scenes.

Aside from [9] and [20] that consider the task of acoustic scene similarity retrieval, much of the work dedicated to acoustic scene modeling is applied to the classification task. An overview of the efforts for the latter task is provided by the IEEE Audio and Acoustic Signal Processing (AASP) Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [26], held in 2013 to compare different approaches to scene classification.

The methods that performed the best all followed a similar structure, where low-level feature representations, such as MFCCs, were calculated over short windows and fed separately to a classifier, such as an SVM, for training. During the classification stage, the recording to be classified was similarly processed into short windows represented by feature vectors which were classified individually by the classifier and the classifier output was combined over the recording through majority vote. This last step is known as late integration and allows the classifier to work on individual, shorter-scale sound events instead of trying to classify the recording as a whole. By doing this, the classifier is able to identify and exploit the distinct events that characterize the scene instead of trying to model its average structure. This is in agreement with the psychoacoustic results related to the perception of auditory scenes presented above.

III. Wavelet scattering

Scattering transforms are time-shift invariant representations of audio signals computed by applying auditory and modulation 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 having a high discriminative power.

A. 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 original class memberships. To discard this superfluous source of variability, signals are mapped into a time-shifting invariant feature space before training the classifier, eliminating the need for this classifier to explicitly learn this invariance and increasing the robustness of its predictions.

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.

B. 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 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 $(J_1 \times Q_1 - 1)$, 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 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 sequel, 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.

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 the Heisenberg theoretical limit of optimal 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 (IFFT) 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 $\boldsymbol{x}$ over scales of duration T, with spectral resolution varying with constant Q. In this way, they are closely related to the mel-frequency spectrogram and related features, such as MFCCs.

C. 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 a local time-shifting invariant representation $\mathbf{S_1}\boldsymbol{x}(t, \gamma_1)$, they can be recovered by a second wavelet transform followed by a pointwise modulus operator.

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 and so have center frequencies $2^{-\gamma_2/Q_2}\xi_2$ for γ_2 taking values between 0 and $(J_2 \times Q_2 - 1)$. 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 sequel, 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 $\boldsymbol{x_2}$ to guarantee local invariance to time-shifting, which yields

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

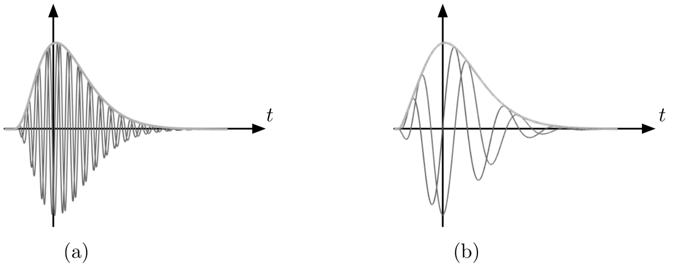


Fig. 1. Gammatone wavelets $\psi(t)$ in the time domain with quality factors (a) $Q = 4$ and (b) $Q = 1$. Oscillations represent the real and imaginary parts. The envelope represents the complex modulus.

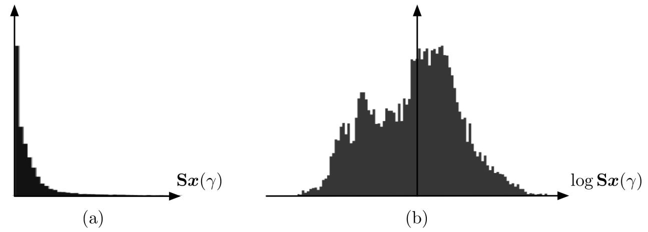


Fig. 2. Histogram of values taken by the first-order scattering coefficient $\mathbf{S}\boldsymbol{x}(\gamma)$, corresponding to a center acoustic frequency of 302 Hz, (a) before and (b) after logarithmic compression.

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.

D. Gammatone wavelets

Wavelets $\boldsymbol{\psi}_{\boldsymbol{\gamma_1}}(t)$ and $\boldsymbol{\psi}_{\boldsymbol{\gamma_2}}(t)$ are designed as fourth-order Gammatone wavelets with one vanishing moment [27], and are shown in Figure 1. In the context of auditory scene analysis, the asymmetric envelope of Gammatone wavelets is more biologically plausible than the symmetric, Gaussian-like envelope 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 [28]. 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 [29]. 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.

IV. Feature design

Prior to constructing models (for classification or similarity estimation), it is beneficial to process scattering coefficients to improve invariance, normality, and generalization power. In this section, we review three transformations which achieve these properties, namely logarithmic compression and standardization.

A. Logarithmic compression

Many algorithms in pattern recognition, including nearest-neighbor classifiers and support vector machines (SVMs), tend to work best when all features follow a standard normal distribution across all training instances [30]. 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.

For the task of musical genre recognition, second-order scattering coefficients $\mathbf{S_2}\boldsymbol{x}(t,\gamma_1,\gamma_2)$ are sometimes normalized by the corresponding first-order scattering coefficients $\mathbf{S_1}\boldsymbol{x}(t,\gamma_1)$, since this decorrelates them from one another [15]. We note that taking the logarithm of such renormalized coefficients yields

$$\log \frac{\mathbf{S_2}\boldsymbol{x}(t,\gamma_1,\gamma_2)}{\mathbf{S_1}\boldsymbol{x}(t,\gamma_1)} = \log \mathbf{S_2}\boldsymbol{x}(t,\gamma_1,\gamma_2) - \log \mathbf{S_1}\boldsymbol{x}(t,\gamma_1), \quad (6)$$

i.e. a linear combination of the logarithms of first- and second-order coefficients. As such, a nonlinear renormalization becomes a linear transformation, which can be learned by a linearly discriminative classifier.

B. Standardization

Let $\mathbf{S}\boldsymbol{x}(\gamma,n)$ be a dataset, where the γ and n respectively denotes feature and sample indices, respectively. It has been found that SVMs should be trained on scaled features with null mean and unit variance so as to avoid mismatch in numeric ranges [30]. To standardize $\mathbf{S}\boldsymbol{x}(\gamma,n)$, 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 training set only, and the same affine transformation is then applied to all samples in both the training and the test sets.

V. Supervised classification system

We are now ready to present the supervised method that we will apply to the acoustic scene classification task using scattering transforms. It consists of two parts, a classifier and a temporal integration method. While the classifier is an SVM, we shall consider two integration approaches, early and late.

A. Support vector machine classifier

/ support vector machine (SVM) is a binary discriminative classifier that attempts to find a separating hyperplane between a pair of classes of that provides the maximal margin between the two sets [31]. From another point of view, it attempts to find a direction on which to project in order to maximize separation between two point clouds. Since it is not a generative, but discriminative model, it is able to pick out those feature vectors that best characterize a particular class of auditory scenes compared to another. This property, coupled with its ease of use and efficient implementations [32] have rendered SVMs a popular choice for many classification systems.

At the classification stage, the SVM gives a class label as output, which is one of the two classes it models. For multiclass settings, a one-vs.-one approach is often used, where an SVM is trained for each pair of classes. During classification, the outputs of all pairwise SVMs is computed and the class selected the largest number of times is chosen.

Finally, the SVM can also be extended to handle nonlinear decision surfaces. This is obtained using the "kernel trick", which exploits the fact that all calculations in the SVM rely only on the inner product between vectors. By replacing this standard inner product by some other positive definite symmetric mapping, such as a Gaussian radial basis function (RBF), this is equivalent to performing a linear SVM after mapping the feature vectors into some high-dimensional space. A non-linear mapping then results in a non-linear decision surface in the orignal space.

B. Early vs. late temporal integration

Owing to the scarcity of salient events in many natural scenes, fine-grained classification is only made possible by integrating signal information over a long temporal context. Indeed, whereas a few seconds is often sufficient to recognize a speaker, a musical instrument, or a genre, up to 30 seconds may be required to disambiguate two classes of auditory scenes with overlapping semantic content, e.g. a train from a subway station or a quiet street from a park. Depending on whether aggregation is performed in feature space or in decision space (the output of a classifier) the corresponding method is referred to as early or late integration.

A straightforward application of early integration consists in summarizing the multivariate time series of scattering coefficients over the full duration of the auditory scene by computing their average values over time. In the definition of the scattering transform (see Section III), this is equivalent to increasing the support T of the low-pass filter $\boldsymbol{\phi}(t)$ up to infinity. With a slight abuse of notation, we denote the summarized features by

$$\mathbf{S}\boldsymbol{x}(\gamma) = \int_{-\infty}^{+\infty} \mathbf{S}\boldsymbol{x}(t,\gamma)\,\mathrm{d}t. \tag{7}$$

Conversely, a late integration scheme relies on probabilistic assignments $\mathbb{P}\left[y \,|\, \mathbf{S}\boldsymbol{x}(t,\gamma)\right]$ over short-term windows of length T obtained from some supervised model, such as a classifier. These are subsequently aggregated to produce a final decision

$$\hat{y} = \arg\max_{y} \rho\Big(\big\{\mathbb{P}\left[y \,|\, \mathbf{S}\boldsymbol{x}(t,\gamma)\right]\big\}_t\Big), \tag{8}$$

where $\hat{y}$ is the estimated class label and ρ is a reduction function, such as sum, product, or majority vote [33].

A potential drawback of early integration is that it reduces a sequence of feature vectors to a single average. If the underlying signal is static in nature, this may be advantageous, as the temporal variation of the feature vectors constitutes an unwanted source of variability. For many audio signals, and in particular auditory scenes, this is very far from the truth, and prevents us from identifying the distinct sound events that characterize the signal as these may be drowned out by the background during the averaging step. In this case, a late integration approach would perform better, since it operates on the individual feature vectors and only the classification results are aggregated over time. Using experiments in acoustic scene classification below, we will demonstrate that the late integration approach does indeed succeed over early integration when scattering coefficients are used.

VI. Acoustic scene similarity retrieval

As discussed in Section II, 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 (see Section V-B), this is not directly feasible in the unsupervised case. As the detection of events is still an open problem [34], 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. Letting μ_m^u denote the centroid of the cluster c_m^u, we attempt to minimize the following objective function:

$$J(C_u) = \sum_{m} \sum_{x_l^u \in c_m^u} \|x_l^u - \mu_m^u\|^2. \tag{9}$$

This is known as k-means clustering. 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 [35], 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 [36]:

$$K_{mn}^{uv} = \exp\left(-\frac{\|\mu_m^u - \mu_n^v\|^2}{\|\mu_m^u - \mu_{m,q}^u\|\|\mu_n^v - \mu_{n,q}^v\|}\right). \tag{10}$$

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 then 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
$$\max_{m,n} K_{mn}^{uv}, \tag{11}$$
- 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
$$\frac{1}{M^2}\sum_{m,n} K_{mn}^{uv} \tag{12}$$
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 centroid is weighted according to the number of frames belonging to its cluster. Each scene s_u is thus described by a signature p_u of M clusters, where $p_u = \{(\mu_1^u, w_1^u), (\mu_2^u, w_2^u), \ldots, (\mu_M^u, w_M^u)\}$) and μ_m^u and w_m^u are the centroid and the weight of the mth cluster, respectively. 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 [37]. 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 times the "ground distance" moved in terms of the histogram bins. In our case, the histogram counts are the clusters weights w_m^u defined over the bins formed by the centroids μ_m^u, so the ground distance is the distance over these cluster centroids.

To compute the $\widehat{\text{EMD}}$, we use the implementation proposed in [38]. Given two signatures p_u and p_v, the $\widehat{\text{EMD}}$ is computed by solving the following linear program:

$$\begin{aligned}\widehat{\text{EMD}}(p_u, p_v) = &\left(\min_{\{f_{nm}\}} \sum_{n,m} f_{nm} D_{nm}^{uv}\right) \\ &+ \left|\sum_n w_n^u - \sum_m w_m^v\right| \max_{n,m}\{D_{nm}^{uv}\}.\end{aligned} \tag{13}$$

$$\text{s.t.} \quad f_{nm} \geq 0 \quad \sum_m f_{nm} \leq w_n^u \quad \sum_n f_{nm} \leq w_m^v$$

$$\sum_{n,m} f_{nm} = \min\left(\sum_n w_n^u, \sum_m w_m^v\right)$$

where $\{f_{nm}\}$ is the flow between the cluster weights w_n^u and w_m^v, that is, the amount transported from the n^{th} bin to "supply the demand" of the m^{th} bin. We denote by D^{uv} the ground distance, a matrix containing the pairwise distances between the centroids sets μ^u and μ^v which is computed from K:

$$D_{mn}^{uv} = 1 - K_{mn}^{uv}.$$

To get the final similarity measure between the scenes s_u and s_v, an extended Gaussian kernel K^s is computed:

$$K_{uv}^s = \exp\left(-\frac{\widehat{\text{EMD}}(p_u, p_v)}{A}\right) \tag{14}$$

with A a scaling parameter, set to the mean value of the $\widehat{\text{EMD}}$ between all the scenes. The resulting kernel K^s is known as an EMD kernel, and it should be noted that there is no guarantee that this kernel is positive definite.

VII. Experiments

The implementations of the presented methods and the experimental protocol are available online.[1]

A. Datasets

The experiments in this paper are carried out on the DCASE 2013 [34] and DCASE 2016 [39] datasets. The DCASE 2013 dataset consists of two parts, namely a public and a private subset, each made up of 100 30-second recordings of various acoustic scenes, sampled at 44100 Hz. The 100 recordings are evenly divided into 10 acoustic scene classes. To build the DCASE 2013 dataset, three different recordists visited a wide variety of locations in Greater London over a period of several months and in each scene recorded a few minutes of audio. 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 consequence, DCASE 2013 dataset enjoys an interesting intra-class diversity while remaining manageable in terms of size, making it suitable for extensive evaluation of algorithmic design choices [20]. In addition, it is still a challenging dataset, with a state-of-the-art system based using handcrafted features as input to SVMs achieving 76% [40] (winner of the DCASE2013 challenge), while recent approaches based on label tree embedding achieve between 84% and 87% on DCASE 2013 depending on the prior knowledge used during training [41].

The public part of the dataset can used for optimizing the acoustic scene classification system and the private part used for computing the resulting accuracy using a five-fold cross validation scheme. For this paper, we shall perform our experiments on the private part of the dataset. The folds used are the same as the ones used during the challenge.

The DCASE 2016 dataset has a similar organization, with 15 classes of acoustic scenes, and is larger than the DCASE 2013 private dataset by one order of magnitude [39].

[1] https://github.com/mathieulagrange/taslp16

TABLE I
CLASSIFICATION ACCURACIES ON THE DCASE 2013 DATASET WITH VARIOUS SETTINGS OF FEATURES, TEMPORAL INTEGRATION STRATEGIES, AND SUPPORT VECTOR MACHINE KERNELS.

	MFCCs	scattering	log-scattering
early, linear	54 ± 19	62 ± 6	66 ± 11
early, RBF	47 ± 16	61 ± 9	63 ± 6
late, linear	60 ± 9	70 ± 8	75 ± 5
late, RBF	68 ± 14	73 ± 6	$\mathbf{78} \pm 6$

TABLE II
CLASS-WISE ACCURACIES ON THE DCASE 2016 DATASET.

Features	MFCCs	log-scattering	log-scattering
Subset	development	development	evaluation
beach	58.3 ± 28.2	83.5 ± 7.1	80.8
bus	58.2 ± 25.4	88.8 ± 11.4	92.3
cafe/restaurant	77.4 ± 14.7	64.5 ± 8.3	50.0
car	92.3 ± 7.7	94.9 ± 6.1	96.2
city_center	96.4 ± 3.8	91.7 ± 9.4	84.6
forest_path	67.7 ± 30.8	93.8 ± 7.8	96.2
grocery_store	44.7 ± 24.8	90.9 ± 6.9	84.6
home	44.8 ± 15.3	56.2 ± 23.9	80.8
library	54.6 ± 20.3	82.0 ± 13.0	65.4
metro_station	94.7 ± 6.8	96.0 ± 2.3	96.2
office	95.7 ± 7.5	87.5 ± 21.7	100
park	52.4 ± 10.4	75.4 ± 7.4	65.4
residential_area	53.2 ± 17.0	44.2 ± 15.0	69.2
train	44.5 ± 23.1	58.3 ± 10.0	53.8
tram	52.1 ± 11.9	83.7 ± 12.2	96.2
Average	65.8 ± 19.3	79.4 ± 15.6	80.8 ± 16.4

B. 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 $\phi(t)$ of duration $T = 372\,\mathrm{ms}$, for a total of $24\,\mathrm{s}$. 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 IV-A).

MFCCs are computed for windows of $50\,\mathrm{ms}$ and hops of $25\,\mathrm{ms}$, 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 $250\,\mathrm{ms}$ long non-overlapping windows so that each window represents structures of similar scale to the scattering coefficients.

C. Experiment #1: acoustic scene classification

Our first task is to compare the performance of the standard scattering transform, the logarithmically compressed scattering (log-scattering) transform, and baseline MFCCs. Secondly, we evaluate the two temporal integration approaches presented above, *early* and *late*.

The different systems are evaluated on the private part of the DCASE 2013 dataset using five-fold stratified cross validation. The folds are identical to those used in the original challenge. The slack parameter C of the SVM is here set to 1. Both linear and Gaussian RBF kernels are used. For the latter, a local scaling method is applied [36]. All feature vectors, that is scattering frames and MFCCs, are standardized prior to be given to the classifier. The standardization is computed with respect to the train/test splits of the cross validation scheme (see Section IV-B). The optimal scaling parameter q of the RBF kernels is learned on the public part of the DCASE 2013 dataset, using five-fold stratified cross validation. For these experiments, both *early* and *late* integration strategies are evaluated (see Section V-B).

D. Experiment #2: acoustic scene similarity retrieval

Having validated the underlying assumptions for the cluster-based auditory scene representation, we evaluate its performance for the task of acoustic scene similarity retrieval.

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. The $p@k$ 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*.

The scaling parameter q of the RBF kernels (see Eq. 10) is set to 10 % of the number of data points to cluster, being the number of scenes for *early* and the number of cluster for RbQ approaches. The clustering for the *RbQ* approaches is done using plain k-means with random initialization and 200 replications. For each method, the numbers of clusters is set to the value among 8, 16 and 32 that leads to the best $p@9$ on the training set.

VIII. RESULTS

The results on acoustic scene classification illustrate the superior performance of the log-scattering transform coupled with late integration, confirming the ideas motivating the construction of our cluster-based auditory scene representation. Using this representation for auditory scene similarity retrieval, we obtain results that improve significantly over traditional BOF and summary approaches, both with log-scattering and MFCCs features.

A. Experiment #1: acoustic scene classification

We report experiments on acoustic scene classification, following the original methodology of the DCASE 2013 challenge, and extend our approach to its 2016 edition.

Role of logarithmic compression: On the DCASE 2013 dataset, logarithmic compression provides a small, albeit consistent boost to acoustic scene classification. Indeed, as discussed in subsection IV-A, and observed in Figure 2, the statistical assumption of normality for scattering coefficients is only approximately satisfied once mapped to a logarithmic, decibel-like scale. This transformation improves the discriminative power of the SVMs trained on scattering coefficients, regardless of the chosen temporal integration strategy (early or late) and the inner product kernel (linear or Gaussian RBF).

MFCCs vs. scattering: As shown on Table I, scattering coefficients outperform MFCCs on the DCASE 2013 private dataset when applying early and integration. This is in accordance with previous results on similarity retrieval, as well as comparisons between mel-frequency coefficients and scattering coefficients for tasks ranging from musical genre recognition and phone segment classification [15] to environmental sound classification [43].

While the scattering coefficients have higher dimension compared to the MFCCs, this does not necessarily imply higher performance. Indeed, if the sounds to be classified were completely classified by their spectral envelopes, MFCCs would be sufficient and may even outperform the scattering transforms, since in this case the latter would contain superfluous modulation information. The results above, however, show the opposite tendency, where this additional information does in fact improve results.

In order to evaluate the generality of our findings, we have taken part in the acoustic scene classification task of the 2016 edition of the DCASE challenge, which was evaluated on a much larger dataset compared to DCASE 2013. For this reason, training support vector machines with the Gaussian RBF kernel proved computationally too demanding and we instead used a linear kernel. Results on DCASE 2013 (see Table I) suggests that this provides comparable, albeit slightly worse results. The results on DCASE 2016 (see Table II).

We compare our results with baseline classifier using linear SVMs and MFCC features, which achieves an average accuracy of 65.5%. With the scattering transform, however, the average accuracy increases to 79.4%. This accuracy increases slightly to 80.8% when calculated on the held-out evaluation dataset.[2] The small change in overall accuracy and relatively stable class-wise accuracies suggest that there is little overfitting in the model, an indication of good generalization power.

Early vs. late integration: On the DCASE 2013 dataset, late integration, i.e. building a prediction by majority voting over 128 local windows of duration $T = 372\,\mathrm{ms}$, outperforms early integration, i.e. classifying the acoustic scene with a global window of duration $T = 24\,\mathrm{s}$. This suggests that there is relevant variability in scales beyond the window size $T = 372$ ms that is not captured by averaging the scattering transform in time (early integration). One potential solution to this is to increase the scale of the second-order modulation wavelets from $J_2 = 12$ octaves to some larger value. This does not result in any performance gain, however (results not shown). What does help is the late integration, since it allows the SVM to successfully classify the few short extracts that characterize the scene which are then used to classify the whole scene through the majority vote. Since the scattering transform, even with larger values of J_2, is not able to separate out these short, discriminative sound events, delegating this task to the late-integrated results in a more powerful classification system.

B. Experiment #2: acoustic scene similarity retrieval

This section presents evaluation results for the acoustic scene similarity retrieval task. The $p@k$ for different settings are shown in Figure 3, illustrating the effect of the different similarity metrics. The metrics is shown for values of k up to 9 which corresponds to the number of elements minus one within each class of scene of the dataset. This allows us to also evaluate the recall capabilities of the system under evaluation.

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 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 [20]. To take advantage of an 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 than *RbQ-w*. This last observation suggests that weighting a centroid according to the number of 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 shared a lot of similar sound sources with only a few sources discriminating between them.

Considering the $p@5$ as in [9] and [20] as our metric, the use of the log-scattering transform versus 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.

IX. 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. We demonstrate the usefulness of scattering-based representations for auditory scenes and the value of late integration using experiments in auditory scene classification. 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

[2] http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification

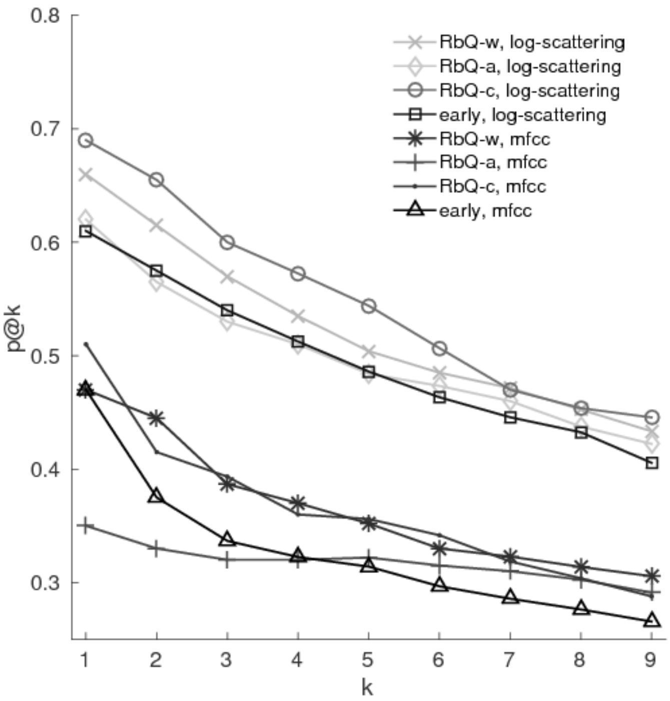


Fig. 3. Acoustic scene similarity retrieval in the DCASE 2013 private dataset: precisions at rank k ($p@k$) obtained for MFCCs and scattering with logarithmic compression, as a function of the rank k.

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].

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] 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.

[10] 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.

[11] I. Nelken. "Processing of complex stimuli and natural scenes in the auditory cortex". In: *Current opinion in neurobiology* 14.4 (2004), pp. 474–480.

[12] C. Guastavino. "The ideal urban soundscape: Investigatng the sound quality of French cities". In: *Acta Acustica United with Acustica* 92 (2006), pp. 945–951.

[13] 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.

[14] B. Logan. "Mel Frequency Cepstral Coefficients for Music Modeling". In: *Proceedings of the International Symposium on Music Information Retrieval*. 2000.

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

[16] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[17] 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.

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

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

[20] 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.

[21] I. Nelken and A. de Cheveigné. "An ear for statistics". In: *Nature neuroscience* 16.4 (2013), pp. 381–382.

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

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

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

[25] 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.

[26] D. Barchiesi et al. "Acoustic scene classification: Classifying environments from the sounds they produce". In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 16–34.

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

[28] H. Fastl and E. Zwicker. *Psychoacoustics: Facts and models, chapter 4*. Springer-Verlag, 2007, pp. 1–463.

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

[30] 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.

[31] C. Cortes and V. Vapnik. "Support-Vector Networks". In: *Machine Learning* 20.3 (1995), pp. 273–297.

[32] C.-C. Chang and C.-J. Lin. "LIBSVM: a library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2.3 (2011), p. 27.

[33] J. Kittler et al. "On combining classifiers". In: *IEEE transactions on pattern analysis and machine intelligence* 20.3 (1998), pp. 226–239.

[34] D. Stowell et al. "Detection and Classification of Acoustic Scenes and Events". In: *IEEE Transactions on Multimedia* 17.10 (2015), pp. 1733–1746.

[35] 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.

[36] 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.

[37] 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.

[38] O. Pele and M. Werman. "Fast and robust earth mover's distances". In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE. 2009, pp. 460–467.

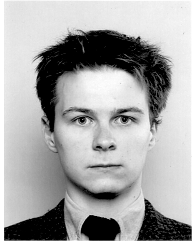
[39] A. Mesaros, T. Heittola, and T. Virtanen. "TUT Database for Acoustic Scene Classification and Sound Event Detection". English. In: *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*. Budapest, Hungary, 2016.

[40] G. Roma et al. "Recurrence quantification analysis features for auditory scene classification". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2013.

[41] H. Phan et al. "Label Tree Embeddings for Acoustic Scene Classification". In: *Proceedings of the 2016 ACM on Multimedia Conference*. 2016, pp. 486–490.

[42] D. Arthur and S. Vassilvitskii. "k-means++: The advantages of careful seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.

[43] J. Salamon and J. P. Bello. "Feature learning with deep scattering for urban sound analysis". In: *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE. 2015, pp. 724–728.



Vincent Lostanlen was born in 1992. He received an engineering degree from TELECOM ParisTech in 2013 and his M.S. degree in acoustics, signal processing, and musical informatics (ATIAM) from the Université Pierre et Marie Curie (UPMC) and the Ircam in Paris, France, in 2013. Since 2013, he is a Ph.D. student at École normale supérieure in Paris, working under the supervision of Stéphane Mallat. His research focuses on defining convolutional operators in the time-frequency domain with applications in audio classification.

Grégoire Lafay was born in 1990. He received the B.S. degree in Acoustic from the University Pierre and Marie Curie (UPMC), Paris, France, and the B.S. degree in Musicology from the Sorbonne University, Paris, France, in 2011. He received his M.S. degree in acoustics, signal processing, and musical informatics (ATIAM) from the UPMC and the Ircam in Paris, France, in 2013. Since 2013, he is a Ph.D. student at IRCCyN, Nantes, France. His research interests include acoustic scene similarity and classification as well as acoustic scene perception.

Joakim Andén received the M.Sc. degree in mathematics from the Université Pierre et Marie Curie, Paris, France, in 2010 and the Ph.D. degree in applied mathematics from Ecole Polytechnique, Palaiseau, France, in 2014. He is currently a postdoctoral researcher with the Program in Applied and Computational mathematics at Princeton University, Princeton, USA. His research interests include signal processing, machine learning, and statistical data analysis. He is a member of the IEEE.

Mathieu Lagrange is a CNRS research scientist at IRCCyN, a French laboratory dedicated to cybernetics. He obtained his Ph.D. in computer science at the University of Bordeaux in 2004, and visited several institutions in Canada (University of Victoria, McGill University) and in France (Orange Labs, TELECOM ParisTech, Ircam). His research focuses on machine listening algorithms applied to the analysis of musical and environmental audio.