Acoustic Scene Classification: Classifying environments from the sounds they produce

Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell and Mark D. Plumbley

Abstract—**In this article we present an account of the state-of-the-art in acoustic scene classification (ASC), the task of classifying environments from the sounds they produce. Starting from a historical review of previous research in this area, we define a general framework for ASC and present different implementations of its components. We then describe a range of different algorithms submitted for a data challenge that was held to provide a general and fair benchmark for ASC techniques. The dataset recorded for this purpose is presented, along with the performance metrics that are used to evaluate the algorithms and statistical significance tests to compare the submitted methods. We use a baseline method that employes MFCCs, GMMs and a maximum likelihood criterion as a benchmark, and only find sufficient evidence to conclude that three algorithms significantly outperform it. We also evaluate the human classification accuracy in performing a similar classification task. The best performing algorithm achieves a mean accuracy that matches the median accuracy obtained by humans, and common pairs of classes are misclassified by both computers and humans. However, all acoustic scenes are correctly classified by at least some individuals, while there are scenes that are misclassified by all algorithms.**

Index Terms—**Machine Listening, Computational Auditory Scene Analysis (CASA), Acoustic Scene Classification, Soundscape Cognition, Computational Auditory Scene Recognition.**

I. Introduction

Enabling devices to make sense of their environment through the analysis of sounds is the main objective of research in *machine listening*, a broad investigation area related to computational auditory scene analysis (CASA)[51]. Machine listening systems perform analogous processing tasks to the human auditory system, and are part of a wider research theme linking fields such as machine learning, robotics and artificial intelligence.

Acoustic scene classification (ASC) refers to the task of associating a semantic label to an audio stream that identifies the environment in which it has been produced. Throughout the literature on ASC, a distinction is made between psychoacoustic/psychological studies aimed at understanding the human cognitive processes that enable our understanding of acoustic scenes [35], and computational algorithms that attempt to automatically perform this task using signal processing and machine learning methods. The perceptual studies have also been referred to as soundscape cognition [15], by defining *soundscapes* as the auditory equivalent of landscapes [43]. In contrast, the computational research has also been called computational auditory scene recognition [38]. This is a particular task that is related to the area of CASA [51], and is especially applied to the study of environmental sounds [18]. It is worth noting that, although many ASC studies are inspired by biological processes, ASC algorithms do not necessarily employ frameworks developed within CASA, and the two research fields do not completely overlap. In this paper we will mainly focus on computational research, though we will also present results obtained from human listening tests for comparison.

Work in ASC has evolved in parallel with several related research problems. For example, methods for the classification of noise sources have been employed for noise monitoring systems [22] or to enhance the performance of speech-processing algorithms [17]. Algorithms for sound source recognition [13] attempt to identify the sources of acoustic events in a recording, and are closely related to event detection and classification techniques. The latter methods are aimed at identifying and labelling temporal regions containing single events of a specific class and have been employed, for example, in surveillance systems [40], elderly assistance [26] and speech analysis through segmentation of acoustic scenes [29]. Furthermore, algorithms for the semantic analysis of audio streams that also rely on the recognition or clustering of sound events have been used for personal archiving [19] and audio segmentation [33] and retrieval [53].

The distinction between event detection and ASC can sometimes appear blurred, for example when considering systems for multimedia indexing and retrieval [9] where the identification of events such as the sound produced by a baseball hitter batting in a run also characterises the general environment *baseball match*. On the other hand, ASC can be employed to enhance the performance of sound event detection [28] by providing prior information about the probability of certain

DB is with Capgemini, UK; DG is with Credit Suisse, London, UK; DS is with Queen Mary University of London, UK; MDP is with University of Surrey, Guildford, UK. The work in this article was carried out while the authors were with the Centre for Digital Music, Queen Mary University of London, Mile End Road, London E1 4NS, UK.

This work was supported by the Centre for Digital Music Platform Grant EP/K009559/1 and a Leadership Fellowship EP/G007144/1 both from the UK Engineering and Physical Sciences Research Council (EPSRC).

Published version: *IEEE Signal Processing Magazine* **32**(3):16-34, May 2015. DOI: 10.1109/MSP.2014.2326181

events. To limit the scope of this paper, we will only detail systems aimed at modelling complex physical environments containing multiple events.

Applications that can specifically benefit from ASC include the design of context-aware services [45], intelligent wearable devices [52], robotics navigation systems [11] and audio archive management [32]. Concrete examples of possible future technologies that could be enabled by ASC include smartphones that continuously sense their surroundings, switching their mode to silent every time we enter a concert hall; assistive technologies such as hearing aids or robotic wheelchairs that adjust their functioning based on the recognition of indoor or outdoor environments; or sound archives that automatically assign metadata to audio files. Moreover, classification could be performed as a preprocessing step to inform algorithms developed for other applications, such as source separation of speech signals from different types of background noise. Although this paper details methods for the analysis of audio signals, it is worth mentioning that to address the above problems acoustic data can be combined with other sources of information such as geo-location, acceleration sensors, collaborative tagging and filtering.

From a purely scientific point of view, ASC represents an interesting problem that both humans and machines are only able to solve to a certain extent. From the outset, semantic labelling of an acoustic scene or soundscape is a task open to different interpretations, as there is not a comprehensive taxonomy encompassing all the possible categories of environments. Researchers generally define a set of categories, record samples from these environments, and treat ASC as a supervised classification problem within a closed universe of possible classes. Furthermore, even within pre-defined categories, the set of acoustic events or qualities characterising a certain environment is generally unbounded, making it difficult to derive rules that unambiguously map acoustic events or features to scenes.

In this paper we offer a tutorial and a survey of the state-of-the-art in ASC. We provide an overview of existing systems, and a framework that can be used to describe their basic components. We evaluate different techniques using signals and performance metrics especially created for an ASC signal processing challenge, and compare algorithmic results to human performance.

II. Background: a history of acoustic scene classification

The first method appearing in the literature to specifically address the ASC problem was proposed by Sawhney and Maes [42] in a 1997 technical report from the MIT Media Lab. The authors recorded a dataset from a set of classes including ‘people’, ‘voices’, ‘subway’, ‘traffic’, and ‘other’. They extracted several features from the audio data using tools borrowed from speech analysis and auditory research, employing recurrent neural networks and a k-nearest neighbour criterion to model the mapping between features and categories, and obtaining an overall classification accuracy of 68%. A year later, researchers from the same institution [12] recorded a continuous audio stream by wearing a microphone while making a few bicycle trips to a supermarket, and then automatically segmented the audio into different scenes (such as ‘home’, ‘street’ and ‘supermarket’). For the classification, they fitted the empirical distribution of features extracted from the audio stream to Hidden Markov Models (HMM).

Meanwhile, research in experimental psychology was focussing on understanding the perceptual processes driving the human ability to categorise and recognise sounds and soundscapes. Ballas [4] found that the speed and accuracy in the recognition of sound events is related to the acoustic nature of the stimuli, how often they occur, and whether they can be associated with a physical cause or a sound stereotype. Peltonen *et. al.* [37] observed that the human recognition of soundscapes is guided by the identification of typical sound events such as human voices or car engine noises, and measured an overall 70% accuracy in the human ability to discern among 25 acoustic scenes. Dubois *et al.* [15] investigated how individuals define their own taxonomy of semantic categories when this is not given a-priori by the experimenter. Finally, Tardieu *et al.* [47] tested both the emergence of semantic classes and the recognition of acoustic scenes within the context of rail stations. They reported that sound sources, human activities and room effects such as reverberation are the elements driving the formation of soundscape classes and the cues employed for recognition when the categories are fixed a-priori.

Influenced by the psychoacoustic/psychological literature that emphasised both local and global characteristics for the recognition of soundscapes, some of the computational systems that built on the early works by researchers at the MIT [42], [12] focussed on modelling the temporal evolution of audio features. Eronen *et al.* [21] employed Mel-frequency cepstral coefficients (MFCCs) to describe the local spectral envelope of audio signals, and Gaussian mixture models (GMMs) to describe their statistical distribution. Then, they trained HMMs to account for the temporal evolution of the GMMs using a discriminative algorithm that exploited knowledge about the categories of training signals. Eronen and co-authors [20] further developed on this work by considering a larger group of features, and by adding a feature transform step to the classification algorithm, obtaining an overall 58% accuracy in the classification of 18 different acoustic scenes.

In the algorithms mentioned so far, each signal belonging to a training set of recordings is generally divided into frames of fixed duration, and a transform is applied to each frame to obtain a sequence of *feature vectors*. The feature vectors derived from each acoustic scene are then employed to train a *statistical model* that summarises the properties of a whole soundscape, or of multiple soundscapes belonging to the same category. Finally, a *decision criterion* is defined to assign unlabelled recordings to the category that best matches the distribution of their features. A more formal definition of an ASC framework will be presented in Section III, and the details of a signal processing challenge we have organised to benchmark ASC methods will be presented in Section IV. Here we complete the historical overview of computational ASC and emphasise their main contributions in light of the components identified above.

A. Features

Several categories of audio features have been employed in ASC systems. Here we present a list of them, providing their rationale in the context of audio analysis for classification.

1) Low-level time-based and frequency-based audio descriptors: several ASC systems [1, GSR][1][20], [34] employ features that can be easily computed from either the signal in the time domain or its Fourier transform. These include (among others) the *zero crossing rate* which measures the average rate of sign changes within a signal, and is related to the main frequency of a monophonic sound; the *spectral centroid*, which measures the centre of mass of the spectrum and it is related to the perception of *brightness* [25]; and the *spectral roll-off* that identifies a frequency above which the magnitude of the spectrum falls below a set threshold.

2) Frequency-band energy features (energy/frequency): this class of features used by various ASC systems [1, NR CHR GSR][20] is computed by integrating the magnitude spectrum or the power spectrum over specified frequency bands. The resulting coefficients measure the amount of energy present within different sub-bands, and can also be expressed as a ratio between the sub-band energy and the total energy to encode the most prominent frequency regions in the signal.

3) Auditory filter banks: A further development of energy/frequency features consists in analysing audio frames through filter banks that mimic the response of the human auditory system. Sawhney and Maes [42] used Gammatone filters for this purpose, Clarkson *et al.* [12] instead computed Mel-scaled filter bank coefficients (MFCS), whereas Patil and Elahili [1, PE] employed a so-called auditory spectrogram.

4) Cepstral features: MFCCS are an example of cepstral features and are perhaps the most popular features used in ASC. They are obtained by computing the discrete cosine transform (DCT) of the logarithm of MFCS. The name *cepstral* is an anagram of *spectral*, and indicates that this class of features is computed by applying a Fourier-related transform to the spectrum of a signal. Cepstral features capture the spectral envelope of a sound, and thus summarise their coarse spectral content.

5) Spatial features: If the soundscape has been recorded using multiple microphones, features can be extracted from the different channels to capture properties of the acoustic scene. In the case of a stereo recording, popular features include the *inter-aural time difference* (ITD) that measures the relative delay occurring between the left and right channels when recording a sound source; and the *inter-aural level difference* (ILD) measuring the amplitude variation between channels. Both ITD and ILD are linked to the position of a sound source in the stereo field. Nogueira *et al.* [1, NR] included spatial features in their ASC system.

6) Voicing features: Whenever the signal is thought to contain harmonic components, a fundamental frequency f_0 or a set of fundamental frequencies can be estimated, and groups of features can be defined to measure properties of these estimates. In the case of ASC, harmonic components might correspond to specific events occurring within the audio scene, and their identification can help discriminate between different scenes. Geiger *et al.* [1, GSR] employed voicing features related to the fundamental frequency of each frame in their system. The method proposed by Krijnders and Holt [1, KH] is based on extracting tone-fit features, a sequence of voicing features derived from a perceptually motivated representation of the audio signals. Firstly, a so-called cochleogram is computed to provide a time-frequency representation of the acoustic scenes that is inspired by the properties of the human cochlea. Then, the *tonalness* of each time-frequency region is evaluated to identify tonal events in the acoustic scenes, resulting in tone-fit feature vectors.

*7) Linear predictive coefficients (*LPCS*):* this class of features have been employed in the analysis of speech signals that are modelled as autoregressive processes. In an autoregressive model, samples of a signal $\boldsymbol{s}$ at a given time instant t are expressed as linear combinations of samples at L previous time instants:

$$\boldsymbol{s}(t) = \sum_{l=1}^{L} \alpha_l \boldsymbol{s}(t-l) + \epsilon(t) \tag{1}$$

where the combination coefficients $\{\alpha_l\}_{l=1}^{L}$ determine the model parameters and ϵ is a residual term. There is a mapping between the value of LPCS and the spectral envelope of the modelled signal [39], therefore α_l encode information regarding the general spectral characteristics of a sound. Eronen *et al.* [20] employed LPC features in their proposed method.

8) Parametric approximation features: autoregressive models are a special case of approximation models where a signal $\boldsymbol{s}$ is expressed as a linear combination of J basis functions from the set $\{\boldsymbol{\phi}_j\}_{j=1}^{J}$

$$\boldsymbol{s}(t) = \sum_{j=1}^{J} \alpha_j \boldsymbol{\phi}_j(t) + \epsilon(t). \tag{2}$$

Whenever the basis functions $\boldsymbol{\phi}_j$ are parametrized by a set of parameters γ_j, features can be defined according to the functions that contribute to the approximation of the signal. For example, Chu *et al.* [10] decompose audio scenes using the Gabor transform, that is a representation where each basis function is parametrized by its frequency f, its time scale u, its time shift τ and its frequency phase θ; so that $\gamma_j = \{f_j, u_j, \tau_j, \theta_j\}$. The set of indexes identifying non-zero coefficients $j^\star = \{j : \alpha_j \neq 0\}$ corresponds to a set of active parameters $\gamma_{j^\star}$ contributing to the approximation of the signal, and encode events in an audio scene that occur at specific time-frequency locations. Patil and Elahili [1, PE] also extract parametric features derived from the 2-dimensional convolution between the auditory spectrogram and 2D Gabor filters.

[1] Here and throughout the paper the notation [1, XXX] is used to cite the extended abstracts submitted for the DCASE challenge described in Section IV. The code XXX (e.g., “GSR”) corresponds to a particular submission to the challenge (see Table I).

9) Unsupervised learning features: The model (2) assumes that a set of basis functions is defined *a priori* to analyse a signal. Alternatively, bases can be learned from the data or from other features already extracted in an unsupervised way. Nam *et al.* [1, NHL] employed a sparse restricted Boltzman machine (SRBM) to adaptively learn features from the MFCCS of the training data. A SRBM is a neural network that has been shown to learn basis functions from input images which resemble the properties of representations built by the visual receptors in the human brain. In the context of ASC, a SRBM adaptively encodes basic properties of the spectrum of the training signals and returns a sequence of features learned from the MFCCS, along with an activation function that is used to determine time segments containing significant acoustic events.

10) Matrix factorisation methods: The goal of matrix factorisation for audio applications is to describe the spectrogram of an acoustic signal as a linear combination of elementary functions that capture typical or salient spectral elements, and are therefore a class of unsupervised learning features. The main intuition that justifies using matrix factorisation for classification is that the signature of events that are important in the recognition of an acoustic scene should be encoded in the elementary functions, leading to discriminative learning. Cauchi [8] employed non-negative matrix factorisation (NMF) and Benetos *et al.* [6] used probabilistic latent component analysis in their proposed algorithms. Note that a matrix factorisation also outputs a set of activation functions which encode the contribution of elementary functions in time, hence modelling the properties of a whole soundscape. Therefore, this class of techniques can be considered to jointly estimate local and global parameters.

11) Image processing features: Rakotomamonjy and Gasso [1, RG] designed an algorithm for ASC whose feature extraction function comprises the following operations. Firstly, the audio signals corresponding to each training scene are processed using a constant-Q transform, which returns frequency representations with logarithmically-spaced frequency bands. Then, 512×512-pixel grayscale images are obtained from the constant-Q representations by interpolating neighbouring time-frequency bins. Finally, features are extracted from the images by computing the matrix of local gradient histograms. This is obtained by dividing the images into local patches, by defining a set of spatial orientation directions, and by counting the occurrence of edges exhibiting each orientation. Note that in this case the vectors of features are not independently extracted from frames, but from time-frequency tiles of the constant-Q transform.

12) Event detection and acoustic unit descriptors: Heittola *et al.* [27] proposed a system for ASC that classifies soundscapes based on a histogram of events detected in a signal. During the training phase, the occurrence of manually annotated events (such as 'car horn', 'applause' or 'basketball') is used to derive models for each scene category. In the test phase, HMMS are employed to identify events within an unlabelled recording, and to define a histogram that is compared to the ones derived from the training data. This system represents an alternative to the common framework that includes features, statistical learning and a decision criterion, in that it essentially performs event detection and ASC at the same time. However, for the purpose of this tutorial, the acoustic events can be thought as high-level features whose statistical properties are described by histograms.

A similar strategy is employed by Chauduri *et al.* [9] to learn acoustic unit descriptors (AUDS) and classify YouTube multimedia data. AUDS are modelled using HMMS, and used to transcribe an audio recording into a sequence of events. The transcriptions are assumed to be generated by N-gram language models whose parameters are trained on different soundscapes categories. The transcriptions of unlabelled recordings during the test phase are thus classified following a maximum likelihood criterion.

B. Feature processing

The features described so far can be further processed to derive new quantities that are used either in place or as an addition to the original features.

1) Feature transforms: This class of methods is used to enhance the discriminative capability of features by processing them through linear or non-linear transforms. Principal component analysis (PCA) is perhaps the most commonly cited example of feature transforms. It learns a set of orthonormal basis that minimise the Euclidean error that results from projecting the features onto subspaces spanned by subsets of the basis set (the principal components), and hence identifies the directions of maximum variance in the dataset. Because of this property, PCA (and the more general independent component analysis (ICA)) have been employed as a dimensionality reduction technique to project high-dimensional features onto lower dimensional subspaces while retaining the maximum possible amount of variance [1, PE][20], [34]. Nogueira *et al.* [1, NR], on the other hand, evaluate a Fisher score to measure how features belonging to the same class are clustered near each other and far apart from features belonging to different classes. A high Fisher score implies that features extracted from different classes are likely to be separable, and it is used to select optimal subsets of features.

2) Time derivatives: For all the quantities computed on local frames, discrete time derivatives between consecutive frames can be included as additional features that identify the time evolution of the properties of an audio scene.

Once features are extracted from audio frames, the next stage of an ASC system generally consists of learning statistical models of the distribution of the features.

C. Statistical models

Statistical models are parametric mathematical models used to summarise the properties of individual audio scenes or whole soundscape categories from the feature vectors. They can be divided into *generative* or *discriminative* methods.

When working with generative models, feature vectors are interpreted as being generated from one of a set of underlying statistical distributions. During the training stage, the parameters of the distributions are optimised based on the statistics of the training data. In the test phase, a decision criterion is defined to determine the most likely model that generated a particular observed example. A simple implementation of this principle is to compute basic statistical properties of the distribution of feature vectors belonging to different categories (such as their mean values), hence obtaining one class *centroid* for each category. The same statistic can be computed for each unlabelled sample that is assumed to be generated according to the distribution with the closest centroid, and is assigned to the corresponding category.

When using a discriminative classifier, on the other hand, features derived from an unlabelled sample are not interpreted as being generated by a class-specific distribution, but are assumed to occupy a class-specific region in the feature space. One of the most popular discriminative classifiers for ASC is the support vector machine (SVM). The model output from an SVM determines a set of hyperplanes that optimally separate features associated to different classes in the training set (according to a maximum-margin criterion). An SVM can only discriminate between two classes. However, when the classification problem includes more than two categories (as is the case of the ASC task presented in this paper), multiple SVMs can be combined to determine a decision criterion that allows to discriminate between Q classes. In the *one versus all* approach, Q SVMs are trained to discriminate between data belonging to one class and data from the remaining $Q-1$ classes. Instead, in the *one versus one* approach $Q(Q-1)/2$ SVMs are trained to classify between all the possible class combinations. In both cases, the decision criterion estimates the class from an unlabelled sample by evaluating the distance between the data and the separating hyperplanes learned by the SVMs.

Discriminative models can be combined with generative ones. For example, one might use the parameters of generative models learned from training data to define a feature space, and then employ an SVM to learn separating hyperplanes. In other words, discriminative classifiers can be used to derive classification criteria from either the feature vectors or from the parameters of their statistical models. In the former case, the overall classification of an acoustic scene must be decided from the classification of individual data frames using, for example, a majority vote.

Different statistical models have been used for computational ASC, and the following list highlights their categories.

1) Descriptive statistics: Several techniques for ASC [1, KH GSR RNH] employ descriptive statistics. This class of methods is used to quantify various aspects of statistical distributions, including moments (such as mean, variance, skewness and kurtosis of a distribution), quantiles and percentiles.

*2) Gaussian mixture models (*GMMs*):* Other methods for ASC [11], [2] employ GMMs, that are generative methods where feature vectors are interpreted as being generated by a multi-modal distribution expressed as a sum of Gaussian distributions. GMMs will be further detailed in Section A where we will present a baseline ASC system used for benchmark.

*3) Hidden Markov Models (*HMMs*):* This class of models are used in several ASC systems [12], [20] to account for the temporal unfolding of events within complex soundscapes. Suppose, for example, that an acoustic scene recorded in an underground train includes an alert sound preceding the sound of the doors closing and the noise of the electric motor moving the carriage to the next station. Features extracted from these three distinct sounds could be modelled using Gaussian densities with different parameters, and the order in which the events normally occur would be encoded in an HMM transition matrix. This contains the transition probability between different states at successive times, that is the probability of each sound occurring after each other. A transition matrix that correctly models the unfolding of events in an underground train would contain large diagonal elements indicating the probability of sounds persisting in time, significant probabilities connecting events that occur after each other (the sound of motors occurring after the sound of doors occurring after the alert sound), and negligible probabilities connecting sounds that occur in the wrong order (for example the doors closing before the alert sound).

*4) Recurrence quantification analysis (*RQA*):* Roma *et al.* [1, RNH] employ RQA to model the temporal unfolding of acoustic events. This technique is used to learn a set of parameters that have been developed to study dynamical systems in the context of chaos theory, and are derived from so-called recurrence plots which capture periodicities in a time series. In the context of ASC, the RQA parameters include: *recurrence* measuring the degree of self-similarity of features within an audio scene; *determinism* which is correlated to sounds periodicities and *laminarity* that captures sounds containing stationary segments. The outputs of the statistical learning function are a set of parameters that model each acoustic scene in the training set. This collection of parameters is then fed to an SVM to define decision boundaries between classes that are used to classify unlabelled signals.

5) i-vector: The system proposed by Elizalde *et al.* [1, ELF] is based on the computation of the *i-vector* [14]. This is a technique originally developed in the speech processing community to address a speaker verification problem, and it is based on modelling a sequence of features using GMMs. In the context of ASC, the i-vector is specifically derived as a function of the parameters of the GMMs learned from MFCCs. It leads to a low-dimensional representation summarising the properties of an acoustic scene, and is input to a generative probabilistic linear discriminant analysis (PLDA) [30].

D. Decision criteria

Decision criteria are functions used to determine the category of an unlabelled sample from its feature vectors and from the statistical model learned from the set of training samples. Decisions criteria are generally dependent on the type of statistical learning methods used, and the following list details how different models are associated to the respective criteria.

1) One-vs-one and one-vs-all: this pair of decision criteria are associated to the output of a multi-class SVM, and are used to map the position of a features vector to a class, as already described in Section II-C.

2) Majority vote: This criterion is used whenever a global classification must be estimated from decisions about single audio frames. Usually, an audio scene is classified according to the most common category assigned to its frames. Alternatively, a weighted majority vote can be employed to vary the importance of different frames. Patil and Elahili [1, PE], for example, assign larger weights to audio frames containing more energy.

3) Nearest neighbour: According to this criterion, a feature vector is assigned to the class associated to the closest vector from the training set (according to a metric, often the Euclidean distance). A generalisation of nearest neighbour is the *k-nearest neighbour* criterion, whereby the k closest vectors are considered and a category is determined according to the most common classification.

4) Maximum likelihood: This criterion is associated with generative models, whereby feature vectors are assigned to the category whose model is most likely to have generated the observed data according to a likelihood probability.

5) Maximum a posteriori: An alternative to maximum likelihood classification is the *maximum a posteriori (*MAP*)* criterion that includes information regarding the marginal likelihood of any given class. For instance, suppose that a GPS system in a mobile device indicates that in the current geographic area some environments are more likely to be encountered than others. This information could be included in an ASC algorithm through a MAP criterion.

E. Meta-algorithms

In the context of supervised classification, meta-algorithms are machine learning techniques designed to reduce the classification error by running multiple instances of a classifier in parallel, each of which uses different parameters or different training data. The results of each classifier are then combined into a global decision.

1) Decision trees and tree-bagger: A decision tree is a set of rules derived from the analysis of features extracted from training signals. It is an alternative to generative and discriminative models because it instead optimises a set of *if/else* conditions about the values of features that leads to a classification output. Li *et al.* [1, LTT] employed a tree-bagger classifier, that is a set of multiple decision trees. A tree-bagger is an example of a classification meta-algorithm that computes multiple so called *weak learners* (classifiers whose accuracy is only assumed to be better than chance) from randomly-sampled copies of the training data following a process called bootstrapping. In the method proposed by Lee *et al.* the ensemble of weak learners are then combined to determine a category for each frame, and in the test phase an overall category is assigned to each acoustic scene based on a majority vote.

2) Normalized compression dissimilarity and random forest: Olivetti [1, OE] adopts a system for ASC that departs from techniques described throughout this paper in favour of a method based on audio compression and random forest. Motivated by the theory of Kolmogorov complexity which measures the shortest binary program that outputs a signal, and that is approximated using compression algorithms, he defines a normalised compression distance between two audio scenes. This is a function of the size in bits of the files obtained by compressing the acoustic scenes using any suitable audio coder. From the set of pairwise distances, a classification is obtained using a random forest, that is a meta-algorithm based on decision trees.

3) Majoriy vote and boosting: The components of a classification algorithm can be themselves thought as parameters subject to optimisation. Thus, a further class of meta-algorithms deals with selecting from or combining multiple classifiers to improve the classification accuracy. Perhaps the simplest implementation of this general idea is to run several classification algorithms in parallel on each test sample and determine the optimal category by majority vote, an approach that will be also used in Section VI of this article. Other more sophisticated methods include *boosting* techniques [44] where the overall classification criterion is a function of linear combinations involving a set of weak learners.

III. A GENERAL FRAMEWORK FOR ASC

Now that we have seen the range of machine learning and signal processing techniques used in the context of ASC, let us define a framework that allows us to distill a few key operators and components. Computational algorithms for ASC are designed to solve a supervised classification problem where a set of M training recordings $\{\boldsymbol{s}_m\}_{m=1}^M$ is provided and associated with corresponding labels $\{c_m\}_{m=1}^M$ that indicate the category to which each soundscape belongs. Let $\{\gamma_q\}_{q=1}^Q$ be a set of labels indicating the members of a universe of Q possible categories. Each label c_m can assume one of the values in this set, and we define a set $\Lambda_q = \{m : c_m = \gamma_q\}$ that identifies the signals belonging to the q-th class. The system learns statistical models from the different classes during an off-line training phase, and uses them to classify unlabelled recordings $\boldsymbol{s}_{\text{new}}$ in the test phase.

Firstly, each of the training signals is divided into short frames. Let D be the length of each frame, $\boldsymbol{s}_{n,m} \in \mathbb{R}^D$ indicates the n-th frame of the m-th signal. Typically, D is chosen so that the frames duration is about 50ms depending on the signal's sampling rate.

Frames in the time domain are not directly employed for classification, but are rather used to extract a sequence of features through a transform $\mathcal{T} : \mathcal{T}(\boldsymbol{s}_{n,m}) = \boldsymbol{x}_{n,m}$, where $\boldsymbol{x}_{n,m} \in \mathbb{R}^K$ indicates a vector of features of dimension K. Often, $K \ll D$ meaning that $\mathcal{T}$ causes a dimensionality reduction. This is aimed at obtaining a coarser representation of the training data where members of the same class result in similar features (yielding *generalisation*), and members of different classes can be distinguished from each other (allowing *discrimination*). Some systems further manipulate the features using feature transforms, such as in the method proposed by Eronen *et al.* [20]. For clarity of notation, we will omit this additional feature processing step from the description of the ASC framework, considering any manipulation of the features to be included in the operator $\mathcal{T}$.

Individual features obtained from time-localised frames cannot summarise the properties of soundscapes that are constituted by a number of different events occurring at different times. For this reason, sequences of features extracted from signals belonging to a given category are used to learn statistical models of that category, abstracting the classes from their empirical realisations. Let $\boldsymbol{x}_{n,\Lambda_q}$ indicate the features extracted from the signals belonging to the q-th category. The function $\mathcal{S} : \mathcal{S}\left(\left\{\boldsymbol{x}_{n,\Lambda_q}\right\}\right) = \mathcal{M}$ learns the parameters of a statistical model $\mathcal{M}$ that describes the global properties of the training data. Note that this formulation of the statistical learning stage (also illustrated in Figure 1) can describe a discriminative function that requires features from the whole training set to compute separation boundaries between classes. In the case of generative learning, the output of the function $\mathcal{S}$ can be separated into Q independent models $\{\mathcal{M}_q\}$ containing parameters for each category, or into M independent models $\{\mathcal{M}_m\}$ corresponding to each training signal.

Once the training phase has been completed, and a model $\mathcal{M}$ has been learned, the transform $\mathcal{T}$ is applied in the test phase to a new unlabelled recording $\boldsymbol{s}_{\text{new}}$, leading to a sequence of features $\boldsymbol{x}_{\text{new}}$. A function $\mathcal{G} : \mathcal{G}(\boldsymbol{x}_{\text{new}}, \mathcal{M}) = c_{\text{new}}$ is then employed to classify the signal, returning a label in the set $\{\gamma_q\}_{q=1}^{Q}$.

Most of the algorithms mentioned in Section II follow the framework depicted in Figure 1, and only differ in their choice of the functions $\mathcal{T}$, $\mathcal{S}$ and $\mathcal{G}$. Some follow a seemingly different strategy, but can still be analysed in light of this framework: for example, matrix factorisations algorithms like the one proposed by Benetos *et al.* [6] can be interpreted as combining features extraction and statistical modelling through the unsupervised learning of spectral templates and an activation matrix, as already discussed in Section II-A.

A special case of ASC framework is the so-called *bag-of-frames* approach [2], named in an analogy with the *bag-of-words* technique for text classification whereby documents are described by the distribution of their word occurrences. Bag-of-frames techniques follow the general structure shown in Figure 1, but ignore the ordering of the sequence of features when learning statistical models.

IV. Challenge on Detection and classification of acoustic scenes and events

Despite a rich literature on systems for ASC, the research community has so far lacked a coordinated effort to evaluate and benchmark algorithms that tackle this problem. The challenge on detection and classification of acoustic scenes and events (DCASE) has been organised in partnership with the IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee in order to test and compare algorithms for ASC and for event detection and classification. This initiative is in line with a wider trend in the signal processing community aimed at promoting reproducible research [50]. Similar challenges have been organised in the areas of music information retrieval [36], speech recognition [5] and source separation [46].

A. *The* DCASE *dataset*

Existing algorithms for ASC have been generally tested on datasets that are not publicly available [42], [20], making it difficult if not impossible to produce sustainable and reproducible experiments built on previous research. Creative-commons licensed sounds can be accessed for research purposes on freesound.org[2], a collaborative database that includes environmental sounds along with music, speech and audio effects. However, the different recording conditions and varying quality of the data present in this repository would require a substantial curating effort to identify a set of signals suited for a rigorous and fair evaluation of ASC systems. On the other hand, the adoption of commercially available databases such as the Series 6000 General Sound Effects Library[3] would constitute a barrier to research reproducibility due to their purchase cost.

The DCASE challenge dataset [23] was especially created to provide researchers with a standardised set of recordings produced in 10 different urban environments. The soundscapes have been recorded in the London area and include: 'bus', 'busy-street', 'office', 'openairmarket', 'park', 'quiet-street', 'restaurant', 'supermarket', 'tube' (underground railway) and 'tubestation'. Two disjoint datasets were constructed from the same group of recordings each containing ten 30s long clips for each scene, totalling 100 recordings. Of these two datasets, one is publicly available and can be used by researchers to train and test their ASC algorithms; the other has been held-back and has been used to evaluate the methods submitted for the challenge.

B. *List of submissions*

A total of 11 algorithms were proposed for the DCASE challenge on ASC from research institutions worldwide. The respective authors submitted accompanying extended abstracts describing their techniques which can be accessed from the DCASE website [4]. The following table lists the authors and titles of the contributions, and defines acronyms that are used throughout the paper to refer to the algorithms.

In addition to the methods submitted for the challenge, we designed a benchmark baseline system that employes MFCCs, GMMs and a maximum likelihood criterion. We have chosen to use these components because they represent standard practices in audio analysis which are not specifically tailored to the ASC problem, and therefore provide an interesting comparison with more sophisticated techniques.

V. Summary table of algorithms for ASC

Having described the ASC framework in Section III and the methods submitted for the DCASE challenge throughout Section II and in Section IV, we now present a table that summarises the various approaches.

VI. Evaluation of algorithms for ASC

A. *Experimental design*

A system designed for ASC comprises training and test phases. Researchers who participated to the DCASE challenge

[2] http://freesound.org

[3] http://www.sound-ideas.com/sound-effects/series-6000-sound-effects-library.html

[4] http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/

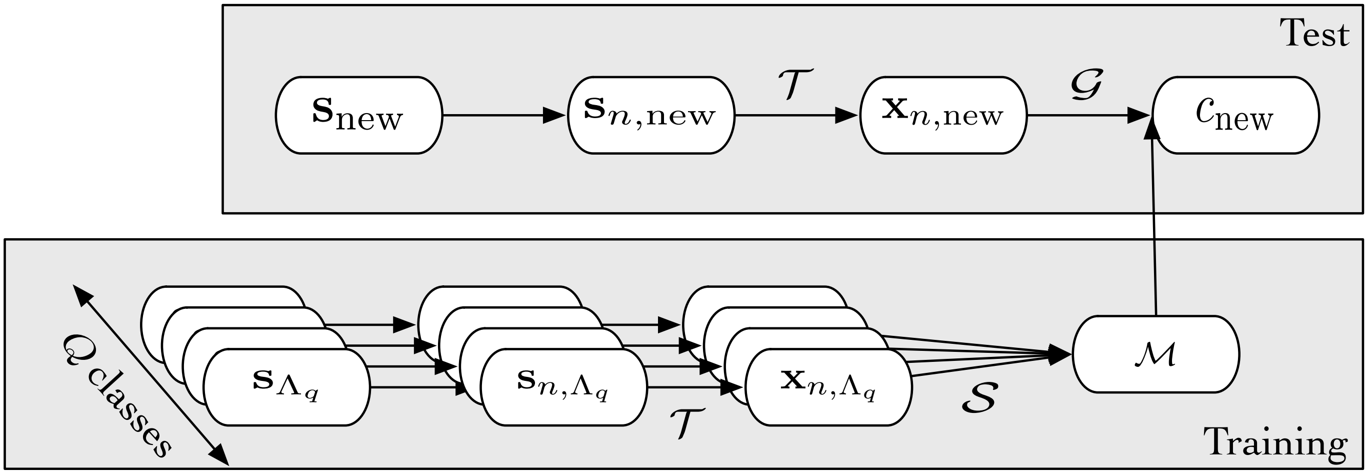

Test
$\mathbf{s}_{\text{new}}$
$\mathbf{s}_{n,\text{new}}$
$\mathcal{T}$
$\mathbf{x}_{n,\text{new}}$
$\mathcal{G}$
c_{new}
Q classes
$\mathbf{s}_{\Lambda_q}$
$\mathbf{s}_{n,\Lambda_q}$
$\mathcal{T}$
$\mathbf{x}_{n,\Lambda_q}$
$\mathcal{S}$
$\mathcal{M}$
Training

Figure 1: Supervised classification framework for acoustic scene classification.

Acronym	**Authors**	**Title**
RNH	G. Roma, W. Nogueira and P. Herrera	Recurrence quantification analysis features for auditory scene classification
RG	A. Rakotomamonjy and G. Gasso	Histogram of gradients of time-frequency representations for audio scene classification
GSR	J. T. Geiger, B. Schuller and G. Rigoll	Recognising acoustic scenes with large-scale audio feature extraction and SVM
CHR	M. Chum, A. Habshush, A. Rahman and C. Sang	IEEE AASP Scene classification challenge using hidden Markov models and frame based classification
NHL	J. Nam, Z. Hyung and K. Lee	Acoustic scene classification using sparse feature learning and selective max-pooling by event detection
NR	W. Nogueira, G. Roma, and P. Herrera	Sound scene identification based on MFCC, binaural features and a support vector machine classifier
PE	K. Patil and M. Elhilali	Multiresolution auditory representations for scene classification
KH	J. Krijnders and G. A. T. Holt	A tone-fit feature representation for scene classification
ELF	B. Elizalde H. Lei, G. Friedland and N. Peters	An I-vector based approach for audio scene detection
LTT[a]	David Li, Jason Tam, and Derek Toub	Auditory scene classification using machine learning techniques
OE	E. Olivetti	The wonders of the normalized compression dissimilarity representation

Table I: List of algorithms submitted for the DCASE challenge on ASC.

[a] The original LTT submission achieved low accuracy due to a bug in a Matlab toolbox - here we are presenting the results obtained with the correct implementation.

were provided with a public dataset that includes ground truth labels indicating the environment in which sounds have been recorded. Training, test and optimisation of design parameters can be performed by partitioning this dataset into training and test subsets, a standard practice in machine learning that is further discussed below. To obtain a fair evaluation reflecting the conditions of a real-world application where sounds and labels are unknown to the algorithms, the methods submitted to the DCASE challenge were tested on a private dataset.

1) Cross-validation: Recall from Figure 1 that statistical models are learned from elements of the training data that belong to different classes, and therefore depend on the particular signals available for training. This represents a general problem of statistical inference occurring every time models are learned using a limited set of data, and is associated with a sampling error or bias. For example, to learn a statistical model of the sounds produced in the office environment, we would ideally need complete and continuous historical recordings from every office in the world. By only analysing data recorded from one or several offices we are bound to learn models that are biased towards the sounds present within the available signals. However, if the training data are rich enough to include sounds produced in most office environments, and if these sounds are effectively modelled,

Method	Features	Statistical model	Decision criterion
Sawhney and Maes [42]	Filter bank	None	Nearest neighbour → majority vote
Clarkson *et al.* [12]	MFCs	HMM	Maximum likelihood
Eronen *et al.* [20]	MFCCs, low-level descriptors, energy/frequency, LPCs → ICA, PCA	Discriminative HMM	Maximum likelihood
Aucouturier [2]	MFCCs	GMMs	Nearest neighbour
Chu *et al.* [10]	MFCCs, parametric (Gabor)	GMMs	Maximum likelihood
Malkin and Waibel [34]	MFCCs, low-level descriptors → PCA	Linear auto-encoder networks	Maximum likelihood
Cauchi [8]	NMF		Maximum likelihood
Benetos [6]	PLCA		Maximum likelihood
Heittola *et al.* [27]	Acoustic events	Histogram	Maximum likelihood
Chaudhuri *et al.* [9]	Acoustic unit descriptors	N-gram language models	Maximum likelihood
DCASE Submissions			
Baseline	MFCCs	GMMs	Maximum likelihood
RNH	MFCCs	RQA, moments → SVM	-
RG	Local gradient histograms (learned on time-frequency patches)	Aggregation → SVM	One versus one
Method	Features	Statistical model	Decision criterion
GSR	MFCCs, energy/frequency, voicing	Moments, percentiles, linear regression coeff. → SVM	Majority vote
CHR	Energy/frequency	SVM	One versus all, majority vote
NHL	Learned (MFCCs → SRBM)	Selective max pooling → SVM	One versus all
NR	MFCCs, energy/frequency, spatial → Fisher feature selection	SVM	Majority vote
PE	filter bank → parametric (Gabor) → PCA	SVM	One versus one, weighted majority vote
KH	Voicing	Moments, percentiles → SVM	-
ELF	MFCCs	i-vector → pLDA	Maximum likelihood
LTT	MFCCs	Ensemble of classification trees	majority vote → treebagger
OE	Size of compressed audio	Compression distance -> ensemble of classification trees	- → random forest

Table II: Summary and categorisation of computational methods for ASC. The acronyms after the author(s) name(s) in the method column are defined in Table I. Arrows indicate sequential processing, for example when statistical parameters learned from features are fed to an SVM to obtain separating hyperplanes. In some cases the decision criterion of SVMs (one versus all, one versus one, or alternative) is not specified in the reference. However, it is always specified when the discriminative learning is performed on frames and an overall classification is determined by a majority vote or a weighted majority vote. Note that for each work cited only the method leading to best classification results have been considered.

then the sampling bias can be bounded, and models can statistically infer general properties of office environments from an incomplete set of measurements. Cross-validation is employed to minimise the sampling bias by optimising the use of a set of available data. The collection of labelled recordings is partitioned into different subsets for training and testing so that all the samples are used in the test phase. Different partition methods have been proposed in the literature for this purpose [7]. To evaluate the algorithms submitted to the DCASE challenge we employed a so-called stratified 5-fold cross-validation of the private dataset. From 100 available recordings, five independent classifications are performed, so that each run contains 80 training recordings and 20 test recordings. The partitions are designed so that the five test subsets are disjoint, thus allowing to perform the classification of each of the 100 signals in the test phases. In addition, the proportion of signals belonging to different classes is kept constant in each training and test subset (8 signals per class in the former and 2 signals per class in the latter) to avoid class biases during the statistical learning.

2) Performance metrics: Performance metrics were calculated from each classification obtained using the training and test subsets, yielding 5 results for each algorithm. Let Γ be the set of correctly classified samples. The classification *accuracy* is defined as the proportion of correctly classified sounds relative to the total number of test samples. The *confusion matrix* is a $Q \times Q$ matrix whose (i, j)-th element indicates the number of elements belonging to the i-th class that have been classified as belonging to the j-th class. In a problem with $Q = 10$ different classes, chance classification has an accuracy of 0.1 and a perfect classifier as an accuracy of 1. The confusion matrix of a perfect classifier is a diagonal matrix whose (i, i)-th elements correspond to the number of samples belonging to the i-th class.

B. Results

Figure 2 depicts the results for the algorithms submitted to the DCASE challenge (see Table I for the acronyms of the methods). The central dots are the percentage accuracies of each technique calculated by averaging the results obtained from the 5 folds, and the bars are the relative confidence intervals. These intervals are defined by assuming that the accuracy value obtained from each fold is a realisation of a Gaussian process whose expectation is the true value of the overall accuracy (that is, the value that we would be able to measure if we evaluated an infinite number of folds). The total length of each bar is the magnitude of a symmetric confidence interval computed as the product of the 95% quantile of a standard normal distribution $q^{0.95}_{\mathcal{N}(0,1)} \approx 3.92$ and the standard error of the accuracy (that is, the ratio between the standard deviation of the accuracies of the folds and the square root of the number of folds $\sigma/\sqrt{5}$). Under the Gaussian assumption, confidence intervals are interpreted as covering with 95% probability the true value of the expectation of the accuracy.

From analysing the plot we can observe that the baseline algorithm achieves a mean accuracy of 55%, and a group of other methods obtain a similar result in the range between 55% and 65%. Four algorithms (GSR, RG, LTT and RNH) approach or exceed a mean accuracy of 70%. OE performs relatively close to chance level and significantly worse than all the other methods. The boxes displaying the results of the paired tests explained in Section VI-C indicate that a number of systems performed significantly better than baseline.

Finally, the method MV indicated in red refers to a majority vote classifier whose output for each test file is the most common category assigned by all other methods. The mean accuracy obtained with this meta-heuristic out-performs all the other techniques, indicating a certain degree of independence between the classification errors committed by the algorithms. In other words, for almost 80% of soundscapes some algorithms make a correct decision, and the algorithms that make an incorrect classification do not all agree on one particular incorrect label. This allows to combine the decisions into a relatively robust meta-classifier. On the other hand, the performance obtained using MV is still far from perfect, suggesting that a number of acoustic scenes are misclassified by most algorithms. Indeed, this can be confirmed by analysing the confusion matrix of the MV solution. As we can see in Figure 3, the class pairs ('park','quietstreet') and ('tube','tubestation') are commonly misclassified by the majority of the algorithms.

To investigate the poor performance of the method OE, we considered the results obtained on the public DCASE dataset, which are not detailed here for the sake of conciseness. OE obtained the highest classification accuracy of all methods, suggesting that it over-fitted the training data by learning models that could not generalise to the test signals.

C. Ranking of algorithms

The ASC performance has been evaluated by computing statistics among different cross-validation folds. However, all the submitted methods have been tested on every file of the same held-back dataset, and this allows us to compare their accuracy on a file-by-file basis. Recall that $\boldsymbol{s}_p$ indicates a signal in the test set. A binary variable X_p can be assigned to each signal and defined so that it takes the value 1 if the file has been correctly classified and 0 if it has been misclassified. Each X_p can be thus interpreted as a realisation of a Bernoulli random process whose average is the mean accuracy of the classifier.

Given two classifiers $\mathcal{C}_1, \mathcal{C}_2$, and the corresponding variables $X_{\mathcal{C}_1,p}$,$X_{\mathcal{C}_2,p}$, a third random variable $Y_p = X_{\mathcal{C}_1,p} - X_{\mathcal{C}_2,p}$ assumes values in the set $\{-1, 0, +1\}$ and indicates the difference in the correct or incorrect classification of $\boldsymbol{s}_p$ by the two classifiers (that is, $Y = -1$ implies that $\mathcal{C}_1$ has misclassified $\boldsymbol{s}$ and $\mathcal{C}_2$ has correctly classified it; $Y = 0$ means that the two methods return equivalently correct or incorrect decisions, and $Y = 1$ implies that $\mathcal{C}_1$ has correctly classified $\boldsymbol{s}$ and $\mathcal{C}_2$ has misclassified it). A *sign test* [24] can be performed to test the hypothesis that the expected value of Y is equal to zero. This is equivalent to performing a paired test evaluating the hypothesis that the performance of the two classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$ is the same. Hence, being able to reject this hypothesis at a fixed probability level provides a method to rank the algorithms.

The grey boxes in Figure 2 represent groups of methods whose accuracy is not significantly different when tested on the DCASE dataset, according to the sign tests ranking criterion evaluated between pairs of different methods. Methods enclosed in the same box cannot be judged to perform better or worse according to the chosen significance level. Starting with the least accurate algorithms, we can observe that the performance of OE is significantly different compared with all the other techniques. Then a clusters of methods ranging from ELF to CHR do not perform significantly differently from the baseline. GSR and RG can be said to have significantly higher accuracy if compared to the baseline method, but not if compared to NR, NHL or CHR. Finally RNH is not significantly more accurate than GSR, RG and LTT, but outperforms all the remaining methods. Note that we do not include the results of the majority vote meta-heuristic in the ranking, as a paired sign test assumes the variables $X_{\mathcal{C}_1,p}, X_{\mathcal{C}_2,p}$ to be statistically independent, and this assumption is violated in the case of MV.

D. Distribution of algorithmic soundscapes classification accuracies

Further analysis of the classification results can be carried out to understand whether there are individual soundscape

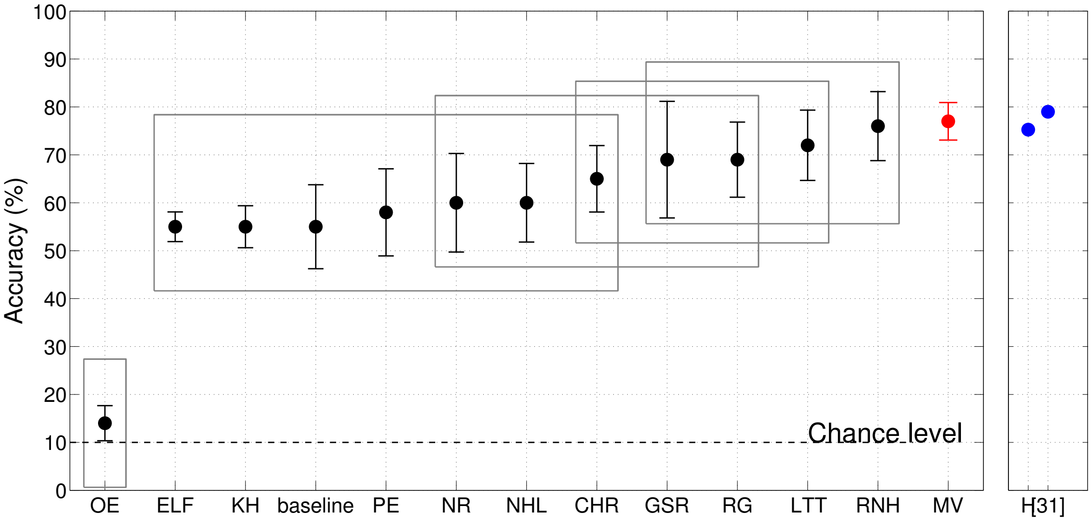


Figure 2: Mean values and confidence intervals of the accuracy of methods for ASC evaluated on the DCASE private dataset using stratified 5-fold cross-validation. The boxes enclose methods that cannot be judged to perform differently with a significance level of 95%. Please see Table I for the definition of the algorithms' acronyms. MV is a majority vote classifier which assigns to an audio recording the label that is most commonly returned by the other methods. 'H' indicates the median human accuracy, as obtained through the test described in Section VII, while '[31]' refers to the human accuracy obtained by Krijnders and Holt. Note that the confidence intervals displayed for the algorithmic results are not directly comparable to the variations in human performance, and hence only the median human performance is depicted. See Figure 6 for more details on the distribution of human accuracies.

recordings in the DCASE dataset that are classified more accurately than others. After evaluating each method with a 5-fold cross-validation, every signal s_p is classified by all the algorithms, resulting in a total of 12 estimated categories. Figure 4 shows a scatter-plot of the mean classification accuracy obtained for each file, and a histogram of the relative distribution. We can observe that some acoustic scenes belonging to the categories 'bus', 'busy-street', 'quietstreet' and 'tubestation' are never correctly classified (those at 0%). In general, the classification accuracy among soundscapes belonging to the same category greatly varies, with the exception of the classes 'office' and 'restaurant' that might contain distinctive events or sound characteristics resulting in more consistent classification accuracies.

E. Pairwise similarity of algorithms decisions

While the results in Figure 2 demonstrate the overall accuracy achieved by algorithms, they do not show which algorithms tend to make the same decisions as each other. For example, if two algorithms use a very similar method, we would expect them to make a similar pattern of mistakes. We can explore this aspect of the algorithms by comparing their decisions pairwise against one another, and using the number of disagreements as a distance measure. We can then visualise this using multidimensional scaling (MDS) to project the points into a low-dimensional space which approximately honours the distance values [16, chapter 10].

Results of MDS are shown in Figure 5. We tested multiple dimensionalities and found that 2D (as shown) yielded a sufficiently low stress to be suitably representative. The OE submission is placed in a corner of the plot, at some distance from the other algorithms; that submission achieved low scores on the private testing data. As a whole, the plot does not appear to cluster together methods by feature type, as MFCC and non-MFCC approaches are interspersed, as are SVM and non-SVM approaches.

VII. Human listening test

In order to determine a human benchmark for the algorithmic results on ASC, we have designed a crowdsourced online listening test in which participants were asked to classify the public DCASE dataset by listening to the audio signals and choosing the environment in which each signal has been recorded from the 10 categories 'bus','busy-street','office','openairmarket','park','quiet-street','restaurant','supermarket','tube' and 'tubestation'.

In designing the listening experiment we chose not to divide the classification into training and test phases because we were interested in evaluating how well humans can recognise the acoustic environments basing their judgement on nothing other than their personal experience. Participants were not presented with labelled training sounds prior to the test, nor were they told their performance during the test.

To maximise the number of people taking the test, we have allowed each participant to classify as many acoustic scenes as he or she liked, while randomising the order in which audio samples appeared in the test to ensure that each file had the same probability to be classified. To avoid potential biases, people who were likely to have worked with the data, and thus likely to know the class labels in advance, did not take the test.

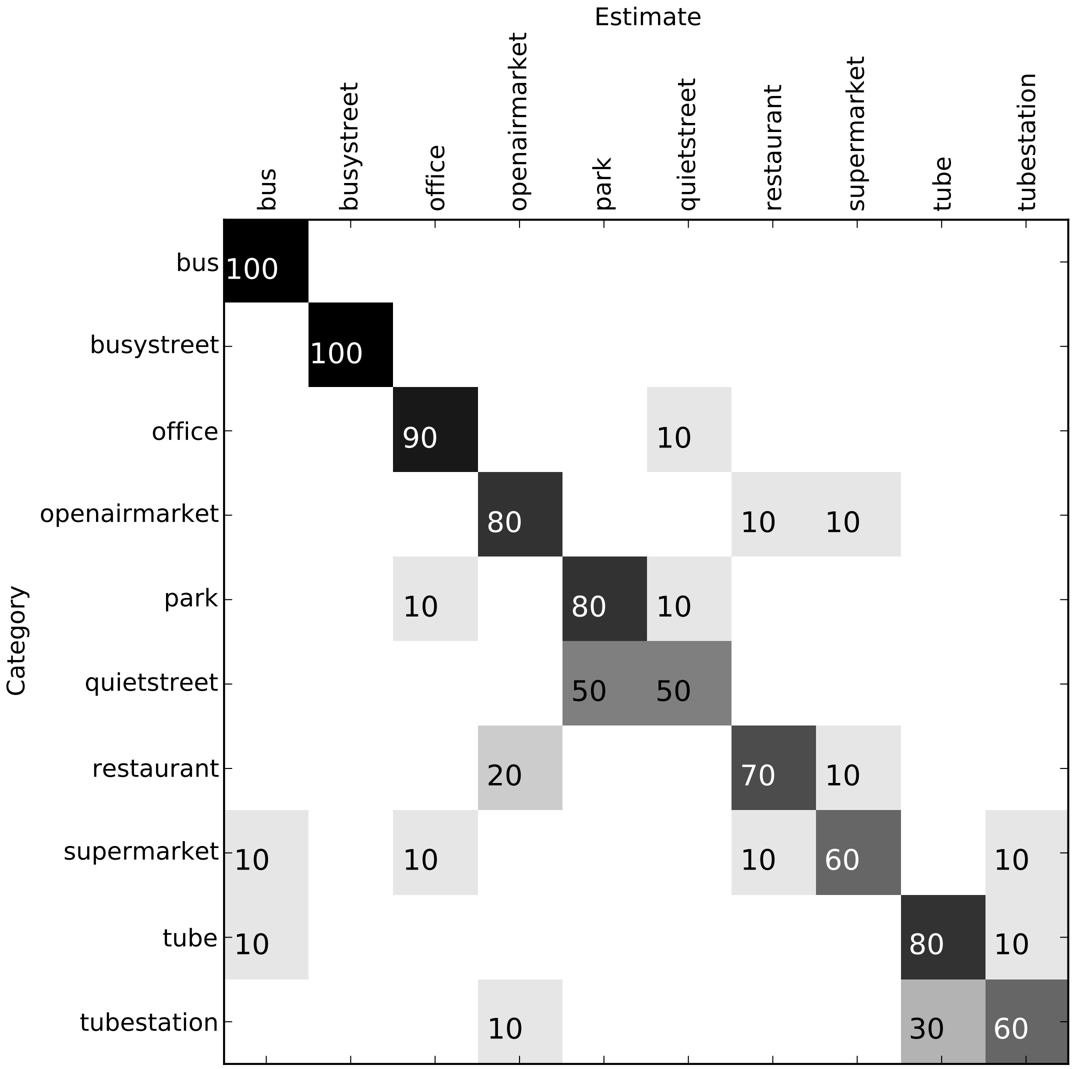


Figure 3: Confusion matrix of MV algorithmic classification results.

A. Human accuracy

A total of 50 participants took part in the test. Their most common age was between 25 and 34 years old, while the most common listening device employed during the test was "high quality headphones". Special care was taken to remove "test" cases or invalid attempts from the sample. This included participants clearly labelled as "test" in the metadata, and participants who only attempted to label only $1-2$ soundscapes, and most of whom achieved scores as low as 0% that points to outliers with a clear lack of motivation. Figure 6 shows that the mean accuracy among all participants was 72%, and the distribution of accuracies reveals that most people scored between 60% and 100%, with two outlier whose accuracy was as low as 20%. Since the distribution of accuracies is not symmetric, we show a box plot summarising its statistics instead of reporting confidence intervals for the mean accuracy. The median value of the participants' accuracy was 75%, the first and third quartiles are located at around 60% and 85%, while the 95% of values lie between around 45% and 100%. Note that, although we decided to include the results from all the participants in the study who classified at least a few soundscapes, the most extreme points (corresponding to individuals who obtained accuracies of about 25% and 100% respectively) only include classifications performed on less than 10 acoustic scenes. Removing from the results participants who achieved about 25% accuracy would result in a mean of 74% a lot closer to the median value. In a more controlled listening test, Krijnders and Holt [31] engaged 37 participants, with each participant asked to listen to 50 public DCASE soundscapes and select one of the 10 categories. The participants were required to listen for the entire duration of the recordings, and use the same listening device. They obtained a mean accuracy of 79%, which is in the same area as the results of our crowdsourced study (75%).

1) Cumulative accuracy: During the test, we asked the participants to indicate their age and the device they used to listen to the audio signals, but we did not observe correlation

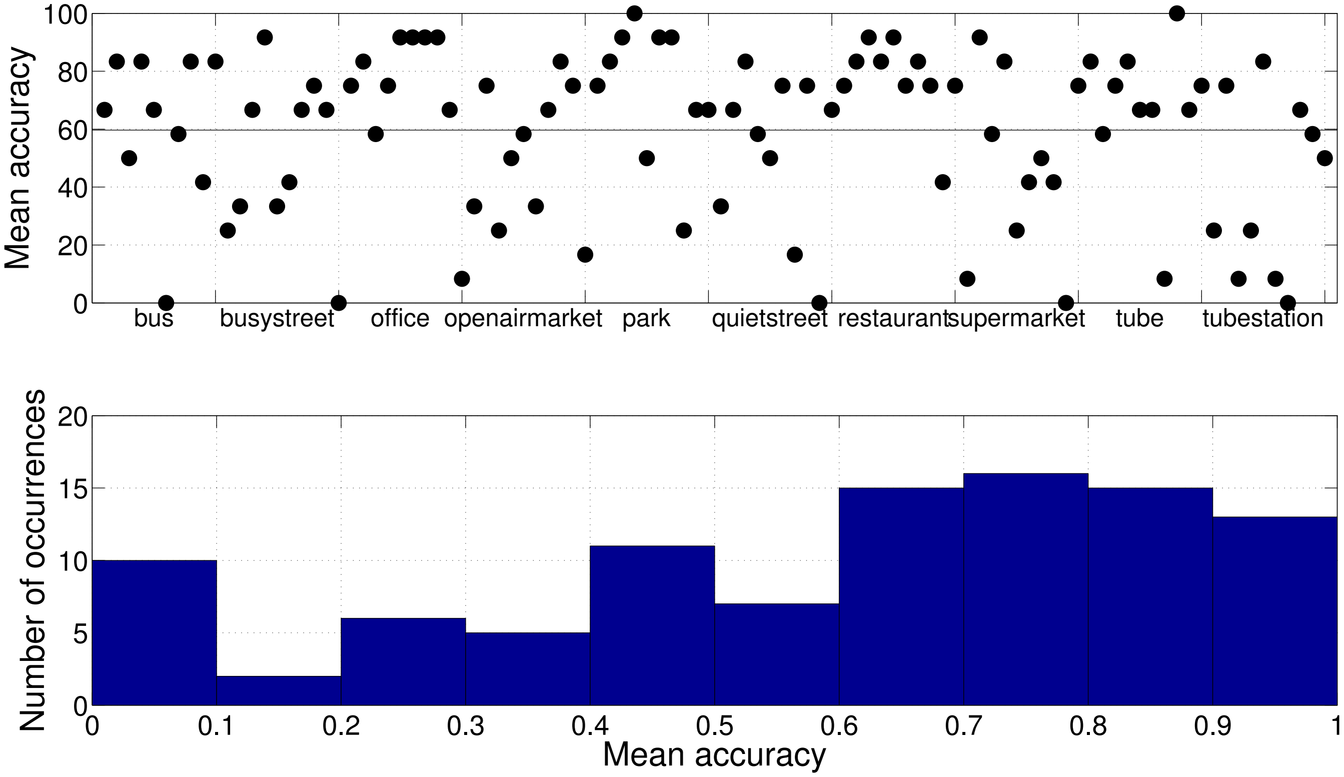


Figure 4: Distribution of algorithmic soundscapes classification accuracies. The solid line in the upper plot represents the average accuracy calculated from all the acoustic scenes. The bottom plot depicts the histogram of mean accuracies resulting from the classification of all 100 soundscapes, highlighting in the left tail that ten soundscapes correctly classified by at most only 10% of the algorithms.

between these variables and the classification accuracy. We did observe a correlation between the number of classified samples and the overall classification accuracy. People who listened to and categorised most or all of the 100 total samples tended to score better than individuals who only classified a few sounds. To assess whether this occurred because participants learned how to better classify the sounds as they progressed in the test, we computed for each individual the cumulative accuracy $\rho(t)$, that is defined as the ratio between the number of correctly classified samples and the total number of classified samples at times $t = 1, \ldots, P$:

$$\rho(t) = \frac{|\Gamma(t)|}{t}. \tag{3}$$

A positive value of the discrete first time derivative of this function $\rho'(t) = \rho(t) - \rho(t-1)$ would indicate that there is an improvement in the cumulative classification accuracy as time progresses. Therefore, we can study the distribution of $\rho'(t)$ to assess the hypothesis that participants have been implicitly training an internal model of the classes as they performed the test. The average of the function $\rho'(t)$ calculated for all the participants results to be -0.0028. A right-tailed t-test rejected with 95% probability that the expectation of $\rho'(t)$ is greater than zero, and a left-tailed t-test failed to reject with the same probability the expectation is smaller that zero, indicating that participants did not improve their accuracy as they progressed through the test. This is a positive finding, as the listening test was designed to avoid training from the exposure to the soundscapes. Having rejected the learning hypothesis, we are left with a selection bias explanation: we believe that people who classified more sounds were simply better able or more motivated to do the test than individuals who found the questions difficult or tedious and did not perform as well.

B. Scenes class confusion matrix

Further insight about the human classification results can be obtained by analysing the overall confusion matrix of the listening test. Figure 7 shows that ‘supermarket’ and ‘openairmarket’ are the most commonly misclassified categories whose samples have been estimated as belonging to various other classes. In addition, there are some common misclassifications between the classes ‘park’ and ‘quietstreet’, and (to a minor extent) between the classes ‘tube’ and ‘tubestation’.

C. Distribution of human soundscapes classification accuracies

To assess if some soundscapes were classified more accurately than others, we conducted a similar analysis for the human performance benchmark to the one described in Section VI-D. Figure 8 depicts the mean accuracy of classification of the 100 soundscapes in the public DCASE dataset, and a

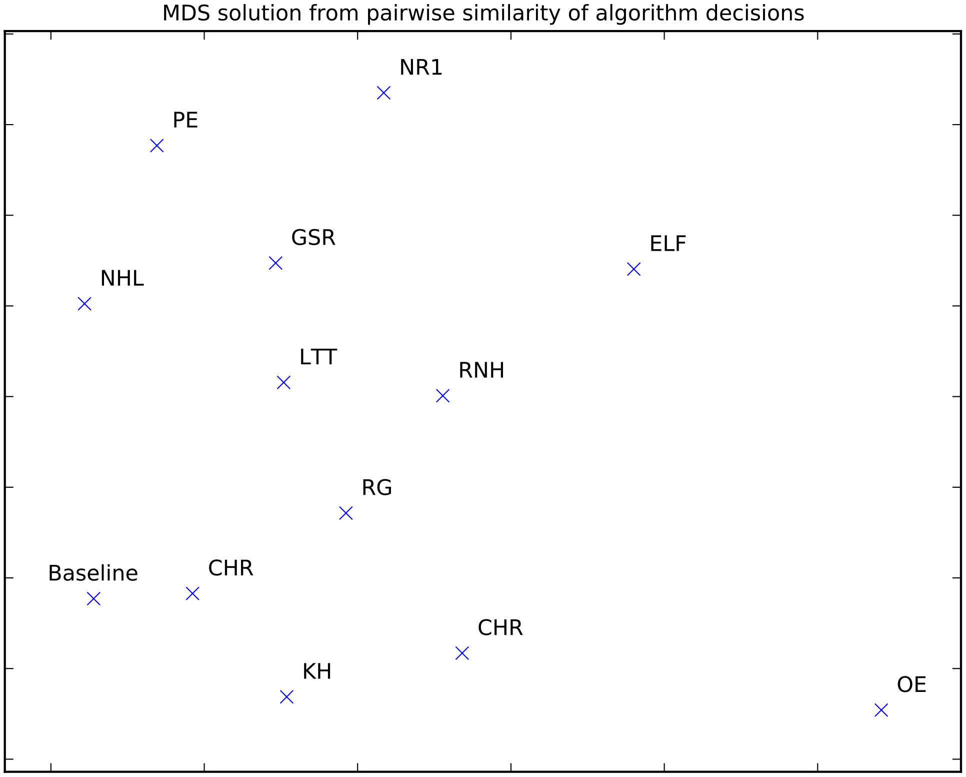


Figure 5: Multidimensional scaling solution (two-dimensional) derived from the pairwise similarities between algorithm labelling decisions. Algorithms which make similar (mis)classifications will tend to appear close to one another. See Section VI-E for details.

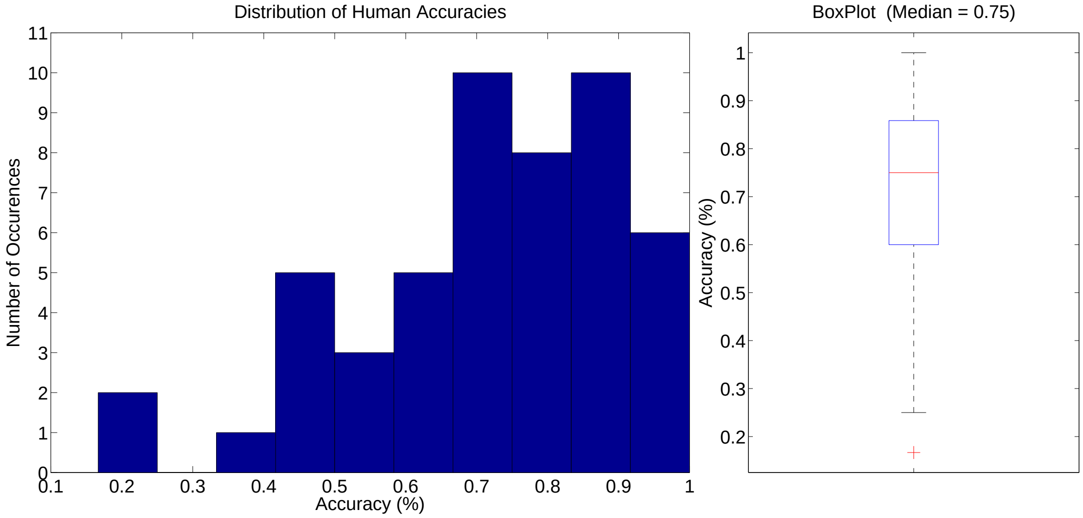
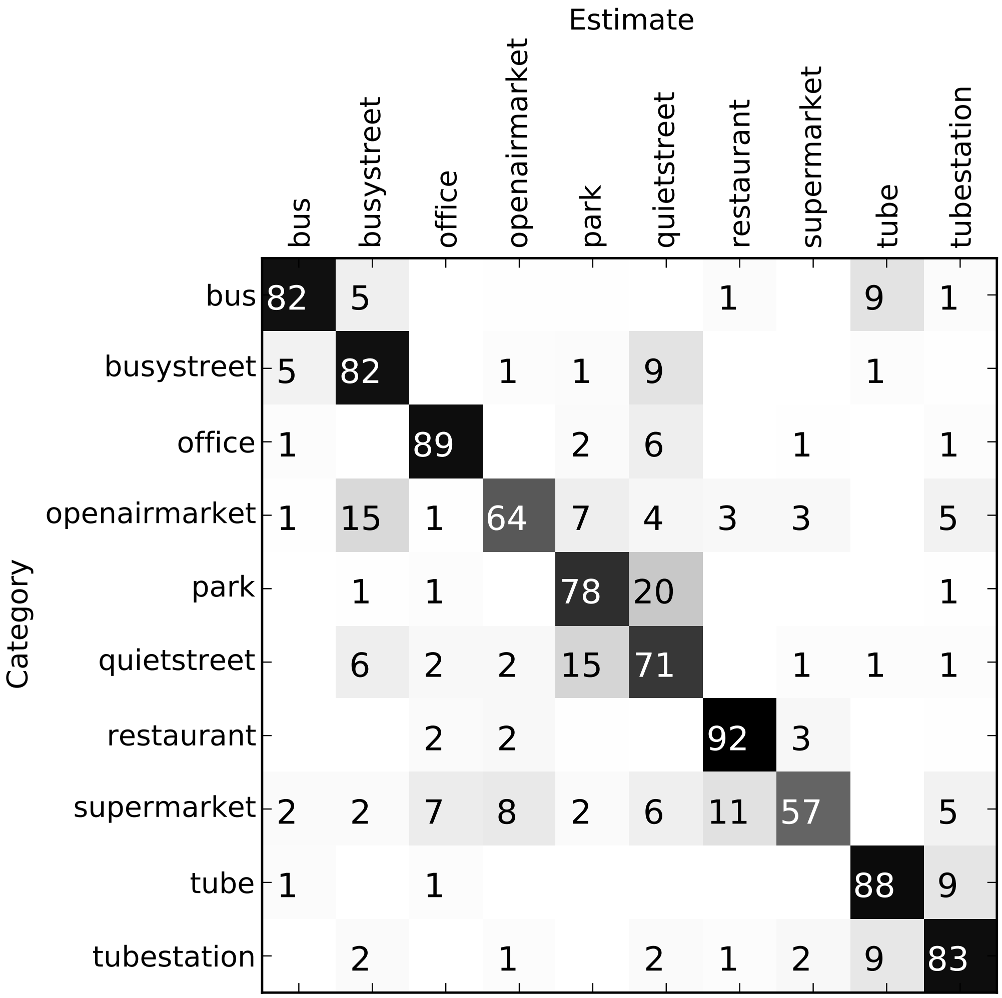


Figure 6: Distribution of human soundscape classification accuracies.

histogram of the relative distribution. The public and private portions of the DCASE dataset are disjoint subsets of the group of recordings produced for the challenge, therefore a paired comparison of the accuracies in Figures 4 and 8 cannot be carried out. Nonetheless, it is informative to compare the trends between the two analysis: it appears that the mean



Category \ Estimate	bus	busystreet	office	openairmarket	park	quietstreet	restaurant	supermarket	tube	tubestation
bus	82	5					1		9	1
busystreet	5	82		1	1	9			1	
office	1		89		2	6		1		1
openairmarket	1	15	1	64	7	4	3	3		5
park		1	1		78	20				1
quietstreet		6	2	2	15	71		1	1	1
restaurant			2	2			92	3		
supermarket	2	2	7	8	2	6	11	57		5
tube	1		1						88	9
tubestation		2		1		2	1	2	9	83

Figure 7: Confusion matrix of human classification results. Note that the rows of the confusion matrix might not add to 100% due to the rounding of percentages

performance for the human classification approaches 80% as opposed to a value around 55% achieved on average by the algorithms. In addition, the distribution of the mean accuracy in the case of a human classification appears more regular, with most soundscapes that are correctly classified most of the times, with only a few outlier scenes whose classification accuracy is below 30%.

VIII. Discussion

By interpreting sophisticated algorithms in terms of a general framework, we have offered a tutorial that uncovers the most important factors to take into account when tackling a difficult machine learning task such as the classification of soundscapes. Inevitably, every abstraction or generalisation is carried out at the expense of omissions in the description of the implementation details of each method. Nonetheless, we think that valuable insights can be gained by analysing the classification results in light of the framework proposed in Section III.

A. Algorithms from the DCASE *challenge*

A first trend regarding the choice of statistical learning function $\mathcal{S}$ can be inferred by analysing the algorithms submitted for the DCASE challenge summarised in Table II. All but one method (ELF) use discriminative learning to map features extracted from the audio signals $\boldsymbol{s}_m$ to class labels c_m. Moreover, most of the algorithms whose mean accuracy is greater or equal than what achieved by the baseline method employ SVM. All techniques that perform significantly better than the baseline except LTT employ a combination of generative and discriminative learning by training an SVM classifier using parameters of models $\mathcal{M}_m$ learned from individual audio scenes. This suggests that models learned from single audio scenes offer an appropriate tradeoff between discrimination and generalisation. On one hand audio signals recorded in the same environment are analysed by learning different statistical models that account for variations between one recording and the next. On the other hand, the parameters of these models occupy localised regions in a parameters space, so that classification boundaries can be learned to discriminate

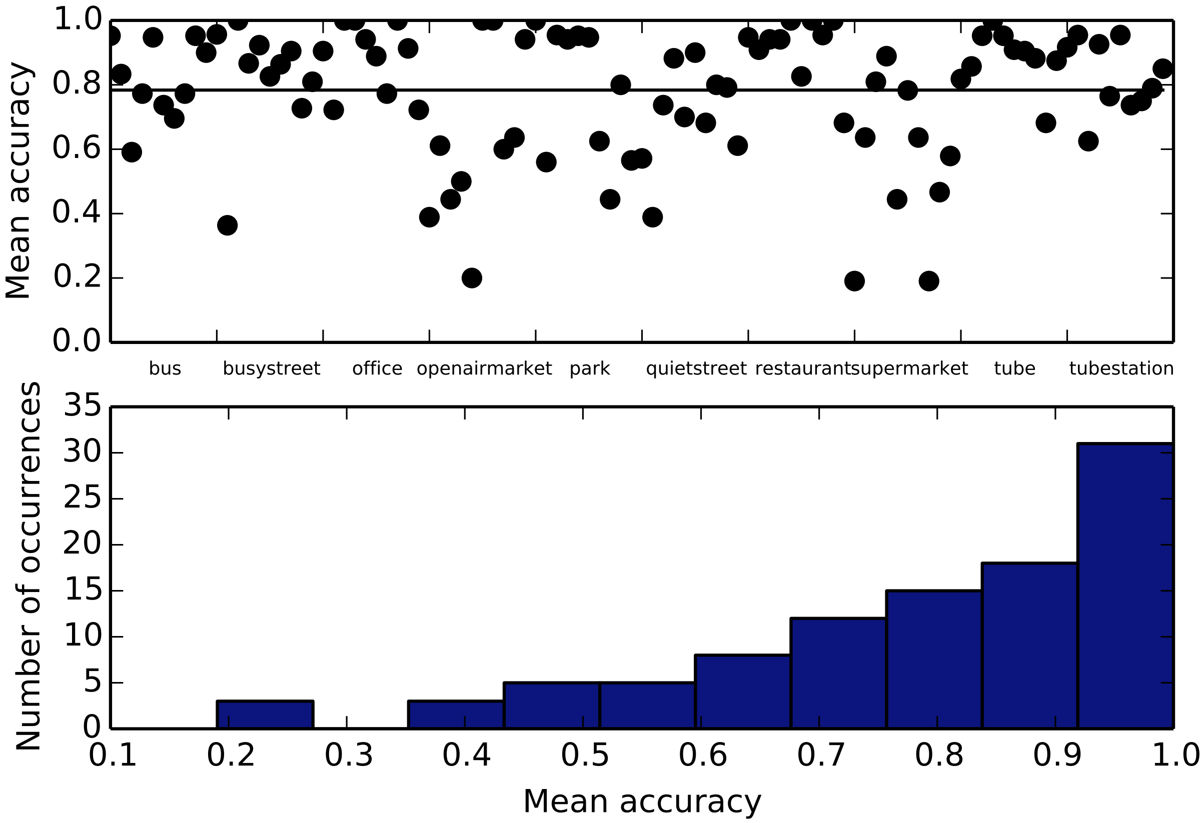


Figure 8: Distribution of human soundscapes classification accuracies. The solid line in the upper plot represents the average accuracy.

between signals recorded in different environments.

A closer analysis of some of the better scoring algorithms (GSR, RG and RNH) reveals a further common design motivation. In different ways, all three methods attempt to model temporal relationships between features extracted from different portions of the signals. RNH employes RQA parameters to encode periodicities (or stationarity) of the MFCC coefficients, RG accounts for time-frequency structures in the audio signals by learning gradient histograms of images derived from their spectrograms, and finally GSR computes linear regression coefficients of local features that encode general trends across a whole scene. This supports the intuitive observation that an ASC method should take into consideration the time evolution of different acoustic events to model complex acoustic scenes.

A further observation derived from analysing Table II is that among the methods that used classification trees in combination with a tree bagger or a random forest algorithm, OE achieved a poor classification performance, while LTT reached the second best mean accuracy. This might suggest that meta-algorithms can be a valuable strategy, but may also be prone to over-fitting.

Finally, a more exploratory remark regards the general use of the framework described in Section III. Aucoutourier [3] studied the performance of a class of algorithms for audio timbre similarity which followed a method similar to the ASC baseline. He reported the existence of a “glass ceiling” as more and more sophisticated algorithms failed to improve the performance obtained using a simple combination of MFCCs and GMMs. To a certain extent, the fact that 7 out of 11 ASC methods did not significantly outperform our baseline might suggest a similar effect, and urges researchers to pursue alternative paradigms. Modelling temporal relationships as described above is one first step in this direction; and perhaps algorithms whose design motivations depart from the ones driving the development of the baseline, such as the normalised compression dissimilarity (OE), might be worth additional investigation.

B. Comparison of human and algorithmic results

When designing the human listening test, we chose to present individuals with samples from the public DCASE dataset to avoid distributing the held-back dataset that was produced to test the algorithms. In addition, we chose not to divide the human task into training and testing phases because we were interested in evaluating how people performed by only drawing from previous experience, and not from prior knowledge about the test set. The different experimental design choices between human and algorithmic experiments do not allow us to perform a statistically rigorous comparison of the classification performances. However, since the public and private DCASE datasets are two parts of a unique session of recordings realised with the same equipment and in the same conditions, we still believe that qualitative comparisons are likely to reflect what the results would have been had we employed a different design strategy that allowed a direct comparison. More importantly, we believe that qualitative conclusions about how well algorithms can approach human capabilities are more interesting than rigorous significance tests on how humans can perform according to protocols (like the 5-fold stratified cross-validation) that are a clearly unnatural task.

Having specified the above disclaimer, several observations can be derived from comparing algorithmic and human classification results. Firstly, Figures 2 and 6 show that RNH achieves a mean accuracy in the classification of soundscapes of the private DCASE dataset that is similar to the median accuracy obtained by humans on the public DCASE dataset. This strongly suggests that the best performing algorithm does achieve similar accuracy compared to a median human benchmark.

Secondly, the analysis of misclassified acoustic scenes summarised in Figures 4 and 8 suggests that, by aggregating the results from all the individuals who took part in the listening test, all the acoustic scenes are correctly classified by at least some individuals, while there are scenes that are misclassified by all algorithms. This observation echoes the problem of *hubs* encountered in music information retrieval, whereby certain songs are always misclassified by algorithms [41]. Moreover, unlike for the algorithmic results, the distribution of human errors shows a gradual decrease in accuracy from the easiest to the most challenging soundscapes. This observation indicates that, in the aggregate, the knowledge acquired by humans through experience still results in a better classification of soundscapes that might be considered to be ambiguous or lacking in highly distinctive elements.

Finally, the comparison of the confusion matrices presented in Figure 3 and Figure 7 reveals that similar pairs of classes ('park' and 'quietstreet', or 'tube' and 'tubestation') are commonly misclassified by both humans and algorithms. Given what we found about the misclassification of single acoustic scene, we do not infer from this observation that the algorithms are using techniques which emulate human audition. An alternative interpretation is rather that some groups of classes are inherently more ambiguous than others because they contain similar sound events. Even if both physical and semantic boundaries between environments can be inherently ambiguous, for the purpose of training a classifier the universe of soundscapes classes should be defined to be mutually exclusive and collectively exhaustive. In other words, it should include all the possible categories relevant to an ASC application, while ensuring that every category is as distinct as possible from all the others.

C. Further research

Several themes that have not been considered in this work may be important depending on particular ASC applications, and are suggested here for further research.

1) Algorithm complexity: A first issue to be considered is the complexity of algorithms designed to learn and classify acoustic scenes. Given that mobile context-aware services are among the most relevant applications of ASC, particular emphasis should be placed in designing methods that can be run with the limited processing power available to smartphones and tablets. The resources-intensive processing of training signals to learn statistical models for classification can be carried out off-line, but the operators $\mathcal{T}$ and $\mathcal{G}$ still need to be applied to unlabelled signals and, depending on the application, might need to be simple enough to allow real-time classification results.

2) Continuous and user-assisted learning: Instead of assuming a fixed set of categories as done in most publications on ASC, a system might be designed to be progressively trained to recognise different environments. In this case, a user should record soundscape examples that are used to train classification models (either on-line or off-line, using the recording device's own computational resources or uploading and processing the signals with remote cloud resources), and progressively add new categories to the system's *memory* of soundscapes. Users could also assist the training by confirming or rejecting the category returned from querying each unlabelled signal, and thus refine the statistical models every time a new classification is performed. Such systems would inevitably require more intervention by the user, but would likely result to be more precise and relevant than totally automated systems.

3) Hierarchical classification: In this paper we have considered a set of categories whose elements are assumed to be mutually exclusive (that is, a soundscape can be classified as 'bus' or 'park' but not both). Alternatively, a hierarchical classification could be considered where certain categories are subsets or supersets of others. For example, a system might be designed to classify between 'outdoor' and 'indoor' environments, and then to distinguish between different subsets of the two general classes. In this context, different costs could be associated with different types of misclassification errors: for example, algorithms could be trained to be very accurate in discriminating between 'outdoor' and 'indoor', and less precise in distinguishing between an outdoor 'park' and an outdoor 'busy-street'.

4) Acoustic scene detection: As a limit case of systems that employes non-uniform misclassification costs, algorithms might be designed to detect a particular environment and group all the other irrelevant categories into an 'others' class. In this case, the system would essentially perform an acoustic scene detection rather than classification.

5) Multi-modal learning: Another avenue of future research consists in fusing multi-modal information to improve the classification accuracy of ASC systems. Video recordings, geo-location information, or temperature and humidity sensors are all examples of data that can be used in conjunction with audio signals to provide machines with context awareness.

6) Event detection and scene classification: The combination of event detection algorithms and ASC which has already been object of research endeavours [27], [9] is likely to benefit from advances in both areas. Information regarding the events occurring in an acoustic scene could be combined with more traditional frame-based approaches to update the probability of categories as different events are detected. For example, while general spectral properties of a soundscape could be used to infer that a signal was likely to have been recorded in either a 'park' or a 'quiet street', detecting the event 'car horn' would help disambiguate between the two. Furthermore, this Bayesian strategy employed to update the posterior probability of different classes could be used to handle transitions between different environments.

7) Testing on different datasets: Finally, datasets that contain sounds from different acoustic environments have been recently released. They include the Diverse Environments Multi-channel Acoustic Noise Database (DEMAND) [48] and the Database of Annotated Real Environmental Sounds (DARES) [49].

IX. Conclusions

In this article we have offered a tutorial in ASC with a particular emphasis on computational algorithms designed to perform this task automatically. By introducing a framework for ASC, we have analysed and compared methods proposed in the literature in terms of their modular components. We have then presented the results of the DCASE challenge, which set the state-of-the-art in computational ASC, and compared the results obtained by algorithms with a baseline method and a human benchmark. On one hand, many of the submitted techniques failed to significantly outperform the baseline system, which was designed to be not optimised for this particular task. On the other hand, some methods significantly out-performed the baseline and approached an accuracy comparable to the human benchmark. Nonetheless, a more careful analysis of the human and algorithmic results highlighted that some acoustic scenes were misclassified by all algorithms, while all soundscapes were correctly classified by at least some individuals. This suggests that there is still scope for improvement before algorithms reach and surpass the human ability to make sense of their environment based on the sounds it produces.

X. Acknowledgments

We would like to thank Dan Ellis, Toumas Virtanen, Jean-Julien Aucouturier, Mathieu Lagrange, Toni Heittola and the anonymous reviewers for having read and commented an early draft of this paper and on our submitted manuscript. Their insights and suggestions have substantially increased the value of this work. We also would like to thank the IEEE AASP Technical Committee for their support in the organisation of the DCASE challenge.

References

[1] Extended abstracts, IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events. Available at http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/, 2013.

[2] 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. *Journal of the Acoustical Society of America*, 112(2):881–891, 2007.

[3] J.-J. Aucouturier and F. Pachet. Improving timbre similarity: How high is the sky? *Journal of negative results in speech and audio sciences*, 1(1), 2004.

[4] J. Ballas. Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 19(2):250–267, 1993.

[5] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green. The PASCAL CHiME speech separation and recognition challenge. *Computer Speech & Language*, 27(3):621–633, 2012.

[6] E. Benetos, M. Lagrange, and S. Dixon. Characterisation of acoustic scenes using a temporally-constrained shift-invariant model. In *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)*, 2012.

[7] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[8] B. Cauchi. Non-negative matrix factorization applied to auditory scene classification. Master's thesis, ATIAM (UPMC / IRCAM / TELECOM ParisTech), 2011.

[9] S. Chaudhuri, M. Harvilla, and B. Raj. Unsupervised learning of acoustic unit descriptors for audio content representation and classification. In *Proceedings of INTERSPEECH*, pages 2265–2268, 2011.

[10] S. Chu, S. Narayanan, and C.-C. Jay Kuo. Environmental sound recognition with time-frequency audio features. *IEEE Trans. on Audio, Speech and Language Processing*, 17(6):1142–1158, Aug. 2009.

[11] S. Chu, S. Narayanan, C.-C. Jay Kuo, and M. J. Matari. Where am I? Scene recognition for mobile robots using audio features. In *IEEE International Conference on Multimedia and Expo*, pages 885–888, 2006.

[12] B. Clarkson, N. Sawhney, and A. Pentland. Auditory context awareness via wearable computing. In *Proceedings Of The 1998 Workshop On Perceptual User Interfaces (PUI'98)*, 1998.

[13] B. Defréville, P. Roy, C. Rosin, and F. Pachet. Automatic recognition of urban sound sources. In *Proceedings of the 120th Audio Engineering Society Convention*, number 6827, 2006.

[14] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel. Support vector machines versus fast scornig in the low-dimensional total variability space for speaker verification. In *Proceedings of INTERSPEECH*, 2009.

[15] D. Dubois, C. Guastavino, and M. Raimbault. A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories. *Acta Acustica united with Acustica*, 92:865–874, 2006.

[16] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2nd edition, 2000.

[17] K. El-Maleh, A. Samouelian, and P. Kabal. Frame level noise classification in mobile environments. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 237–240, 1999.

[18] D. P. W. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, 1996.

[19] D. P. W. Ellis and K. Lee. Minimal-impact audio-based personal archives. In *Proceedings of the Workshop on Continuous Archiving and Recording of Personal Experiences*, pages 39–47, 2004.

[20] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *IEEE Trans. on Audio, Speech and Language Processing*, 14(1):321–329, Jan. 2006.

[21] A. J. Eronen, J. T. Tuomi, A. Klapuri, and S. Fagerlund. Audio-based context awareness - acoustic modeling and perceptual evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 529–532, 2003.

[22] P. Gaunard, C. G. Mubikangiey, C. Couvreur, and V. Fontaine. Automatic classification of environmental noise events by hidden markov models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages 3609–3612, 1998.

[23] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley. A database and challenge for acoustic scene classification and event detection. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2013.

[24] J. D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*. Number 978-1420077612. Chapman and Hall, 5th edition, 2010.

[25] J. M. Grey and J. W. Gordon. Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, 63(5):1493–1500, 1978.

[26] P. Guyot, J. Pinquier, and R. André-Obrecht. Water sounds recognition based on physical models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 793–797, 2013.

[27] T. Heittola, A. Mesaros, A. J. Eronen, and T. Virtanen. Audio context recognition using audio event histogram. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2010.

[28] T. Heittola, A. Mesaros, A. J. Eronen, and T. Virtanen. Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 1:1–13, 2013.

[29] G. Hu and D. Wang. Auditory segmentation based on onset and offset analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):396–405, 2007.

[30] S. Ioffe. Probabilistic linear discriminant analysis. In *Proceedings of the 9th European Conference on Computer Vision (ECCV)*, pages 531–542, 2006.

[31] J. Krijnders and G. A. t. Holt. Tone-fit and MFCC scene classification compared to human recognition. Personal Communication, Oct. 2013.

[32] C. Landone, J. Harrop, and J. Reiss. Enabling access to sound archives through integration, enrichment and retrieval: the easaier project. In *Proceedings of the 8th International Conference on Music Information Retrieval, Vienna, Austria*, September 2007.

[33] L. Lu, H.-J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE Trans. on Audio, Speech and Language Processing*, 10(7):504–516, 2002.

[34] R. G. Malkin and A. Waibel. Classifying user environment for mobile applications using linear autoencoding of ambient audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 509–512, 2005.

[35] S. McAdams. Recognition of sound sources and events. In S. McAdams and E. Bigand, editors, *Thinking in Sound: the Cognitive Psychology of Human Audition*, pages 146—98. Oxford University Press, 1993.

[36] Music Information Retrieval Evaluation eXchange (MIREX). http://music-ir.org/mirexwiki/.

[37] V. T. Peltonen, A. J. Eronen, M. P. Parviainen, and A. P. Klapuri. Recognition of everyday auditory scenes: Potentials, latencies and cues. In *Proceedings of the 110th Audio Engineering Society Convention*, number 5404, 2001.

[38] V. T. Peltonen, J. T. Tuomi, A. Klapuri, J. Huopaniemi, and L. Sorsa. Computational auditory scene recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 1941–1944, 2002.

[39] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Number ISBN: 0-13-285826-6. Prentice-Hall, 1993.

[40] R. Radhakrishnan, A. Divakaran, and P. Smaragdis. Audio analysis for surveillance applications. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 158–161, 2005.

[41] M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

[42] N. Sawhney and P. Maes. Situational awareness from environmental sounds. Technical report, Massachussets Institute of Technology, 1997.

[43] M. Schafer. *The tuning of the world*. Random House, Rochester VT, 1977.

[44] R. E. Schapire. *The boosting approach to machine learning an overview*, volume 171 of *Lecture notes in statistics*, pages 149–172. Springer, 2003.

[45] B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In *Proceedings of the Workshop on Mobile Computing Systems and Applications*, pages 85–90, 1994.

[46] Signal separation evaluation campaign (SiSEC). http://sisec.wiki.irisa.fr/tiki-index.php, 2013.

[47] J. Tardieu, P. Susini, F. Poisson, P. Lazareff, and S. McAdams. Perceptual study of soundscapes in train stations. *Applied Acoustics*, 69:1224–1239, 2008.

[48] J. Thiemann, N. Ito, and E. Vincent. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings. In *21st International Congress on Acoustics*, Montreal, Canada, June 2013. Acoustical Society of America.

[49] M. van Grootel, T. Andringa, and J. Krijnders. DARES-G1: Database of annotated real-world everyday sounds. In *Proceedings of the NAG/DAGA International Conference on Acoustics*, 2009.

[50] P. Vandewalle, J. Kovacevic, and M. Vetterli. Reproducible research in signal processing. *IEEE Signal Processing Magazine*, 26(3):37–47, 2009.

[51] D. Wang and G. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. IEEE Press, 2006.

[52] Y. Xu, W. J. Li, and K. K. Lee. *Intelligent Wearable Interfaces*. ISBN 978-0-470-17927-7. Wiley and Sons, 2008.

[53] T. Zhang and C.-C. Jay Kuo. Audio content analysis for online audiovisual data segmantation and classification. *IEEE Trans. on Audio, Speech and Language Processing*, 9(4):441–457, 2001.

Appendix

A. MFCCs

Mel-frequency cepstral coefficients have been introduced in Section II-A and have been widely used as a feature for audio analysis. Let $\boldsymbol{s}_n \in \mathbb{R}^D$ be a signal frame and $|\hat{\boldsymbol{s}}_n|$ the absolute value of its Fourier transform. The coefficients corresponding to linearly-spaced frequency bins are mapped onto R Mel frequency bands to approximate the human perception of pitches (which can be approximately described as logarithmic, meaning that we are capable of a much better resolution at low frequencies than at high frequencies), resulting in $L \leq D$ coefficients. The magnitude of the Mel coefficients is converted to a logarithmic scale, and the resulting vector is processed using a discrete cosine transform (DCT). Finally, the $K \leq R$ first coefficients are selected and constitute the vector of features $\boldsymbol{x}_n = \mathcal{T}(\boldsymbol{s}_n)$. This last step essentially measures the frequency content of the log-magnitude of the spectrum of a signal, and therefore captures general properties of the spectral envelope. For example, periodic sounds which exhibit spectral peaks at multiples of a fundamental frequency are highly correlated with one or several cosine bases, encoding this information in the value of the corresponding MFCC coefficients. The set of parameters $\theta = \{D, R, K\}$ includes frames dimension, number of Mel bands and number of DCT coefficients which need to be defined when computing the MFCCs. These parameters determine the dimensionality reduction introduced by the features extraction operator, and their choice is governed by the tradeoff between generalisation and discrimination that has been mentioned in Section III.

1) Statistical normalization: To classify features extracted from signals belonging to different categories, it is important to evaluate relative differences between the values of feature vectors belonging to different classes, rather than differences between different coefficients within feature vectors extracted from the same signal. For this reason, during the training phase of the ASC classification algorithm, statistical normalisation is performed as a standard feature processing aimed at avoiding offsets or scaling variations of any of the coefficients within feature vectors. This is accomplished by subtracting the global mean (computed from features extracted from the whole dataset) from each vector $\boldsymbol{x}_{n,m}$, and by dividing each coefficient by the their global standard deviation. After the feature vectors have been normalised, the average and standard deviation of the coefficients $x_{n,m,k}$ are 0 and 1 respectively.

B. GMMs

Gaussian mixture models (GMMS) have been introduced in Section II-C, and are used to infer global statistical properties of the features from local features vectors, which are interpreted as realisations of a generative stochastic process. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^K$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}$, and recall that the notation $\boldsymbol{x}_{n,\Lambda_q}$ identifies features vectors extracted from training signals that belong to the q-th category. Then every such vector is modelled as generated by the following distribution:

$$\boldsymbol{x}_{n,\Lambda_q} \sim \prod_{i=1}^{I} w_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \qquad (4)$$

where I is a fixed number of components, and w_i is a latent variable expressing the probability that a particular observation is generated from the i-th component.

The operator $\mathcal{S}$ takes the collection of features $\boldsymbol{x}_{n,\Lambda_q}$ and learns a global model for the q-th class $\mathcal{M}_q = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{I}$ by estimating the parameters of the Gaussian mixture distribution in Equation (4), which can be accomplished through an expectation-maximisation (EM) algorithm [7]. The only parameter to be set in this case is the number of Gaussian components I which rules a tradeoff between model accuracy and over-fitting. Indeed $\mathcal{S}_I$ must include a sufficient number of components to account for the fact that different events within a soundscape generate sounds with different spectral properties. However, as the number of components becomes too large, the model tends to fit spurious random variations in the training data, hindering the generalisation capabilities of the algorithm when confronted with an unlabelled sound.

C. Maximum likelihood criterion

Once the GMMS $\mathcal{M}_q$ have been learned from the training data, features can be extracted from an unlabelled sound by applying the operator $\mathcal{T}$. The new sequence of features $\boldsymbol{x}_{n,\text{new}}$ is statistically normalised using the same mean and standard deviation values obtained from the training signals, and a likelihood measure $\mathcal{G}$ is employed to evaluate which class is statistically most likely to generate the observed features, hence determining the sound classification. A set of coefficients g_q is computed by evaluating the log-likelihood of the observed data given the model:

$$g_q = p(\boldsymbol{x}_{n,\text{new}} | \mathcal{M}_q) \propto \sum_{i=1}^{I} w_i (\boldsymbol{x}_{n,\text{new}} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i (\boldsymbol{x}_{n,\text{new}} - \boldsymbol{\mu}_i) \qquad (5)$$

and a category is picked based on the most likely model $c^{\star}_{\text{new}} = \arg\min_{q} g_q$.

Note that the baseline system described in this section is an example of a bag-of-frames technique where the ordering of the sequence of features is irrelevant. Indeed, any random permutation of the sequences $\boldsymbol{x}_{n,\Lambda_q}$ does not affect the computation of the GMM parameters, and thus the classification of unlabelled signals.