# A morphological model for simulating acoustic scenes and its application to sound event detection

Grégoire Lafay, Mathieu Lagrange, Mathias Rossignol, Emmanouil Benetos, *Member, IEEE*, and Axel Roebel, *Member, IEEE*

***Abstract*—This paper introduces a model for simulating environmental acoustic scenes that abstracts temporal structures from audio recordings. This model allows us to explicitly control key morphological aspects of the acoustic scene and to isolate their impact on the performance of the system under evaluation. Thus, more information can be gained on the behavior of an evaluated system, providing guidance for further improvements. To demonstrate its potential, this model is employed to evaluate the performance of nine state of the art sound event detection systems submitted to the IEEE DCASE 2013 Challenge. Results indicate that the proposed scheme is able to successfully build datasets useful for evaluating important aspects of the performance of sound event detection systems, such as their robustness to new recording conditions and to varying levels of background audio.**

***Index Terms*—Sound event detection, acoustic scene analysis, computational auditory scene analysis, experimental validation.**

## I. Introduction

OVER the past decades, the amount of recorded audio data documenting our sonic environment has grown considerably. Emerging research areas such as eco-acoustics [1], [2] massively record environmental sounds around the world in order to measure potential animal biodiversity modification over large temporal scales due to human activity or climate change [3]–[5]. Other research areas focus on sound-related human activities for context inference and surveillance [6]–[8].

As part of the aforementioned research areas and applications, the emerging field of *Sound Scene Analysis* (SSA – also called *Acoustic Scene Analysis*) [9] aims to develop approaches and systems for the automatic analysis of environmental sounds and soundscapes (originating from both urban and natural environments). While research methodologies in related fields such as Automatic Speech Recognition (ASR) [10] and Music Information Retrieval (MIR) [11] are now well established, research addressing SSA remains relatively young. Open research questions in this emerging field include: 1) Gaining knowledge about the characteristics of acoustic scenes (e.g. recognizing the recording environment, detecting and seperating sound sources) and how the aforementioned audio scenes can be modelled; 2) Proposing new systems which can be employed in related applications, including eco-acoustics [1], organization of sound collections [12], security / surveillance [7], and smart homes / cities [13]. Being a relatively new research field, only few datasets for SSA are available, though this number may grow as the interest for such problems increases in the scientific and engineering communities; see [14] and [13] for research effort in related sound recognition tasks in urban environments.

Within the field of acoustic scene analysis, the problem of *Sound Event Detection* (SED – also called *Acoustic Event Detection*) focuses on building systems that can automatically detect sound events in an acoustic scene [6], [9], [15]. Typically, a sound event detection system labels temporal regions for each event within an audio recording, resulting in a symbolic description with start and end times, as well as labels for each instance of a specific event type [9]. Problems closely related to SED include automatic speech recognition [10] and automatic music transcription [11] when considering speech and music audio recordings, respectively. The majority of research in SED is directed towards detecting one event at a given time segment (*monophonic sound event detection*), while a smaller subset of research addresses the more challenging problem of detecting overlapping events from audio (*polyphonic sound event detection*) [15]. One major issue in the SED problem is the lack of understanding about the relatively poor performance of state of the art systems [16]. To address this, this paper focuses on building a method for simulating sound scenes that can be used to evaluate sound event detection systems under various event density and background noise conditions. Used in a well-designed experimental protocol, we demonstrate that it allows researchers to efficiently gain insights about the behavior of their algorithms.

This work builds upon the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), which was organised in 2013 by the Center for Digital Music of Queen Mary University of London and by the Institute for Research and Coordination in Acoustics / Music (IRCAM), under the auspices of the Audio and Acoustic Signal Processing (AASP) technical committee of the IEEE Signal Processing Society [16]. The DCASE 2013 Challenge is the second challenge dedicated to this task after the CLEAR Challenge in 2007 [17].

During a research community discussion phase for the setup of the DCASE challenge through a dedicated mailing list,

Grégoire Lafay and Mathieu Lagrange are with the IRCCYN CNRS Laboratory, École Centrale de Nantes, France.

Mathias Rossignol and Axel Roebel are with the IRCAM (CNRS, UPMC), Paris, France.

Emmanouil Benetos is with the Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, London, U.K.

Research partly supported by ANR Houle under reference ANR-11-JS03-005-01. Emmanouil Benetos is supported by a UK Royal Academy of Engineering Research Fellowship (grant no. RF/128).

an interest rose about the potential benefit of considering simulated data in order to carry out a controlled evaluation of submitted systems for the SED task. Varying the power level of the background, the density of the events, their intra class diversity, all seemed important aspects that would be desirable to study, though these would be costly to tackle using recorded and annotated data. To this end, a simulation protocol was needed, which would be based on a model of environmental sound scenes. As discussed in detail in Section VI, we acknowledge that the use of simulated data shall not be considered as sufficient for the final evaluation of engineering systems; that being said, the above described potential benefits are still sufficient to justify pursuing that avenue of research.

Similar simulation approaches have been carried out in more mature fields, for example the CHiME challenge [18]–[20], which focuses on robust ASR using simulated data (in addition to real data); in particular the 3rd CHiME challenge [20] addresses the validity of simulations in the topic of speech enhancement. In addition, Cristoforetti *et al.* [21] developed a simulated multi-microphone multi-language acoustic corpus for distant ASR, as part of the DIRHA research project.

The end goal of the proposed simulation framework is its use for evaluating single-microphone sound event detection systems in the presence of background noise and overlapping sound events (i.e. *polyphonic SED*). Thus, we do not specifically address the problems of (de-)reverberation or distant microphone sound recognition, since these add many other morphological parameters that are outside of the scope of the DCASE challenge [9]. The aforementioned important matters have been addressed in related corpora, such as the 3rd CHiME challenge dataset [20] and the DIRHA corpus [21] for evaluating distant microphone sound recognition, as well as the REVERB challenge dataset [22] for evaluating dereverberation techniques in ASR applications, and are left for future work.

The aim of the model proposed in this paper is to generate sound scenes as a " skeleton of events on a bed of texture ", as coined in [23]. Since the simulated scenes are intended to be analyzed by event detection systems trained on recorded data, our aim is to minimize the discrepancy between the simulated scenes and the recorded ones. Thus, we do not consider approaches based on actual synthesis of sounds. To the authors' knowledge, this is the first system specifically designed to generate sound scenes to be used for the objective evaluation of sound event detection systems. It thus departs significantly from models used in research fields such as wave field synthesis [24], binaural or spatial scene synthesis [25], acoustic event synthesis [26] and texture synthesis [27]–[29].

The proposed model is based on several sequences of sound events emitted by the same source, where each sound event is drawn from a collection of carefully chosen sound samples. The *morphological* aspects of the scene, i.e. which sound sample is present at what time and which level, are then modeled in an abstract manner, allowing us to control high level properties of the scene. The contribution of this paper is threefold: 1) propose a computational model for the generation of simulated datasets of sound scenes, 2) review perceptual considerations that root important morphological aspects of the model, and 3) consider this simulation paradigm to gain knowledge about the behavior of several sound event detection systems developed by different research teams worldwide, initially submitted to the IEEE DCASE challenge [14].

The outline of this paper is as follows: Section II defines the concept of a sound scene and sound collection, drawing from auditory scene analysis theory; Section III presents the proposed model of sound scenes which underlies the simulation process; Sections IV and V present the evaluation framework for event detection systems using simulated sound scenes; then, the use of simulated data to evaluate detection algorithms is discussed in Section VI.

## II. The notion of sound collection

### A. Auditory scene as a sum of sound sources

The proposed model adopts a "source-driven" approach by considering a sound scene as a sum of sound sources. This approach is consistent with the way humans perceive their sonic environment. Studies addressing the Auditory Scene Analysis (ASA) [30] problem, and more specifically the sound segregation process [31] [32] [33] [34], show that humans make sense from their sonic world by isolating information related to individual sound sources. Considering a bottom-up approach, the segregation process relies on generic rules involving Gestalt-like principles [34] to group sounds with similar acoustic indicators (common onset, spectral regularity and harmonicity), as well as similar perceptual attributes (timbre, loudness, perceived location and pitch) into perceptual entities called "auditory streams". Recently, several neurophysiological studies have shown evidence of the existence of auditory streams [35].

Besides ASA studies which mostly consider pure tones or simple complex sounds [32], more recent studies adopting a psycho-linguistic approach to describe recorded sounds have also demonstrated the existence of top-down source-driven grouping processes involved in sound perception. Investigating the qualitative evaluation of urban sound scenes using categorization tasks and linguistic analysis, studies of Dubois and colleagues [36] [37] have shown that listeners categorize sound environments on the basis of semantic features, *i.e.* the meaning attributed to the recalled sound sources.

Considering both the ASA and the psycho-linguistic approach, it seems intuitive for the simulation process to consider separately the sound activity of each sound source of the scene. In practical terms, to materialize these sound activities, each sound source has to be related to a collection of sound recordings. But this approach introduces fundamental questions about the very nature of such a collection. It first questions the existence of a standardized taxonomy of sounds, which should be a hierarchical classification system putting together sounds according to the characteristics they share, where each group is labeled in such a way that a specific name may describe its corresponding class, an instance of it, but also at which level of the classification it fits. Unfortunately, if such taxonomies exist for plants, animals or colors, it is not the case for sounds [38], for two main reasons:

- Sound description and identification are highly subjective. In other words, a same sound may be described quite differently according to the subject. This is due to the relative lack of basic lexicalized terms to describe acoustic phenomena [39].
- Sound description and identification are highly dependent of their context, that is, sound source identification depends on the nature of other co-occurring sound sources [34], [40], [41].

Despite those difficulties, one may take into account some perceptual considerations to guarantee a certain level of ecological validity.

### *B. Action and Sources*

SED tasks evaluate if an algorithm is able to detect a specific set of sound classes. Ideally, to prevent from low generalization capability, the training set of a given class shall be consistent, that is, class exemplars should be representative of the diversity of the sounds suggested by the class label that may occur in the real world. In our case, the class exemplars are the recordings of a sound collection.

Some perceptual considerations can be taken into account to guide the collection building process. First, one may look at the way humans classify / categorize sounds. As explained by [42], " categorization is a cognitive process that unites different entities of an equivalent status ". Among other categorization strategies, several studies show that humans categorize sounds according to 1) the type of source (agent, object, functions) and / or 2) the action / movement causing the sounds [34], [36], [43]–[46].

Human categorization occurs at several levels. Rosch [47] proposes three levels of categorization for real-world objects, namely superordinate, basic, and subordinate. The higher the level, the higher the abstraction degree of the categories. Considering sound perception, Guyot et al. [43] propose a framework where listeners identify sound categories of abstract concepts at a superordinate level (" noise generated by a mechanical excitation "), actions at the basic level (" grating ", " scratching ", " rubbing ") and sources at the subordinate level (" dishes ", " pen " " sharpening ", " door "). Although Houix et al. [42] find some differences by showing that sounds seem " to be categorized as sound sources first and only second as actions ", it appears that source and action are adequate verbal descriptors for category.

One way to make a sound collection consistent is to consider low-level categories as the intra-category diversity decreases with the level. Considering that, one may label a sound collection using a " source-action " couple (e.g. *passing-car*), or at least one of the two, in order to minimize the expected diversity of its recordings. Any name referring to higher category levels may lead to sound collections comprising a too large variety of objects. That definition of a collection raises two issues:

- Building such collections assumes the availability of a large number of recorded sounds to be representative of the diversity suggested by the collection label.
- Adopting a data-centered approach, such collections may lead to a misinterpretation of the results of a detection task for someone who did not build them, as the nature of the entities suggested by the collection labels are ambiguous (*e.g.* a sound collection of *traffic sounds vs.* a sound collection of *passing-car*).

Thus, considering the source-action couple is not sufficient. Generic labels must also be defined for each couple. To do so, one may refer to the works of Gaver [48], who proposes a phenomenological taxonomy of everyday sounds, Niessen *et al.* [38], who assesses the consensus of categories mentioned in 166 papers of different research domain using linguistic analysis, and recently Salamon *et al.* [13], who build a taxonomy of urban sounds based on the work of Brown et al. [49].

Thus, labeling a class using the source-action nomenclature helps us reduce the expected intra-class diversity. However, it does not address the issue of inter-class diversity. Indeed, a naive source-driven approach implies to record in a source-wise way all the sound activities that may occur in an environment. Considering dense environments such as cities or forests, this may raise important practical issues. To circumvent this problem, one may assume that sources do not carry the same potential information, and some are not required to be recorded separately.

### *C. Texture* vs. *Event*

The human brain can easily distinguish between a voice sound and a background of other competing sounds [33]. Considering the example of an urban sound scene, global traffic hubbub sounds are typically uninformative, compared with closer human sounds [50].

Maffiolo [51] shows the existence of two distinct cognitive processes depending on the listener's ability to identify separate sound events. By asking subjects to categorize recordings of urban environments and using linguistic analysis of the verbal descriptions of the categories, she exhibits two cognitive categories of sound environments called respectively " event sequences " and " amorphous sequences ". Event sequences (in which distinct events or sequences of events can be identified) are processed analytically, that is, based on the meaning of the identified sound sources, whereas amorphous sequences (in which no event can be isolated) are processed holistically using global acoustical indicators (intensity, spectral content). This distinction is validated by Guastavino in [39]: using semantic analysis of verbal descriptions of specific sounds populating the urban environment, Guastavino shows that verbal descriptions of low pitched sounds may be divided into two categories called " source events " (sound events which can be attributed to a sound source), and " background noise " (where no identifiable event can be isolated).

Thus, sound perception highly depends on semantic features (source identification), but also on the quantity of " information " carried by the source. Sound sources that carry information of interest are processed independently, whereas the others are processed together, *i.e.* merged in a single stream. Following this, another common distinction is made between two perceptual objects called “ sound events ” and “ sound textures ”. Based on previous studies on vision, McDermott and Simoncelli [28], [52] show that the perception of sound textures may derive from simple statistics of early auditory representations; which would be sufficient to recognize sounds having some temporal homogeneity.

There are also a few formal definitions concerning the texture object [27]. The most notable attempt is made by Saint-Arnaud [53] and Saint-Arnaud and Popat [54]. Following their conclusions, a texture may be understood as a composite object with two hierarchical levels, the top level being the high level pattern, and the leaf level being the atom. The nature of an atom remains adaptable as the latter may be considered at several time scales. Thus and to some extent, a texture may be considered as a concatenation of recordings, each of them being a sequence of atoms. These recordings must comprise at least the high level pattern of the texture, that is, if we consider a texture of ‘gallop’, recordings of sequences of atoms must be at least composed of the first three sounds of hooves.

To summarize the previous statements, it appears that all sounds are not processed as a sum of distinct events:

- amorphous sequences that convey low semantic information are processed holistically;
- sound textures with stable acoustic properties over long period are processed using summary statistics of these acoustics properties.

To circumvent the issue of recording a representative number of sound collections to simulate a sound scene, one can take into account those considerations, and use recordings of mixed sound sources, provided that they can be considered as amorphous sequences or textures. We believe that there exists some links between the notions of amorphous sequences and textures. Both trigger holistic processing based on global acoustical properties for amorphous sequences [36], [51] and summary statistics for textures [52], and both convey a low information content [53]. Amorphous sequences are described as “ background sounds ” with no identifiable events, whereas the texture definition comprises sequences of events such as “ gallop ” that do not meet this last criterion. Considering that, one can consider an amorphous sequence to be a texture, as the physical characteristics of an amorphous sequence remain stable over time, but the reverse is not systematic.

## III. Proposed Model

### A. Model components

From the considerations discussed in Section II, we derive two types of sound collections to be used as basic elements by the proposed simulation process: the “ event collections ” and the “ texture collections ”. For both collections, a stream is modeled as being a temporal sequence of sound recordings issued from the same sound collection. For the texture collection, each recording is an sequence of atoms, or more precisely, a sequence of sound events which follow a periodic or a stochastic pattern. The nature of the sequence to be recorded depends on the considered type of texture. For a texture with a periodic pattern such as gallop, recordings are sequences of events comprising at least the first three sounds of hooves, while for a texture with a stochastic pattern such as “ rain ”, the recordings are simply samples of rain sounds. This method offers some flexibility, as it makes it possible to quickly generate various versions of a same texture with few recorded samples, by varying the apparition order of the sequence. Obviously for a texture to be realistic, sequences have to come from the same recording session. Moreover, as the human brain is very sensitive to repetitions of identical sounds, even when they are individual chunks of white noise [55], a sequence shall not be concatenated with itself or repeated without a sufficient time delay.

To summarize, the proposed source-driven model uses collections as basic elements for the simulation process:

- Each collection is a group of sound recordings.
- Insofar as possible, the label of the collection should be of the form “ source + action ” where source and action labels must be generic.
- There are two types of collections, called respectively the event collections, and the texture collections.
- Sound recordings of a same event collection come from the same sound source.
- Sound recordings of a same texture collection are atomic sequences emitted by one or a mixture of sound sources.
- Sound recordings of a texture collection must at least comprise the high-level pattern of the texture (*e.g.* three sounds of hooves for the gallop texture).
- A texture built from the concatenation of recordings must convey a low semantic information and / or have stable acoustic properties over time.

### B. Design choices

Building on the above discussed matters, the proposed simulation process considers a sound scene as a sum of sound sources. Each sound source activity is a sequence of recorded sound samples emitted by the considered sound source. To generate each track, the model takes into account the following set of four parameters:

1) the mean / variance of the Event to Background power Ratio (EBR) of sound samples
2) the mean / variance of the time interval between consecutive onsets of sound samples
3) the mean / variance of the duration between sound samples
4) the start / end times of the track

As explained in Section II-C, sound events and texture are likely to be treated differently by the human auditory system. Thus, the model explicitly distinguishes them and processes them separately. A track of events is made of discrete sound samples, whereas a texture track consists of one continuous sound, or a seamless concatenation of samples. Thus, for the texture track, the mean/variance time interval between samples as well as the variance EBR are set to 0.

Each semantic track, texture or event, is related to a specific sound collection. As discussed in Section II, a sound collection may be seen as a group of similar sound recordings, each of which comprise sound signals that are emitted by the same

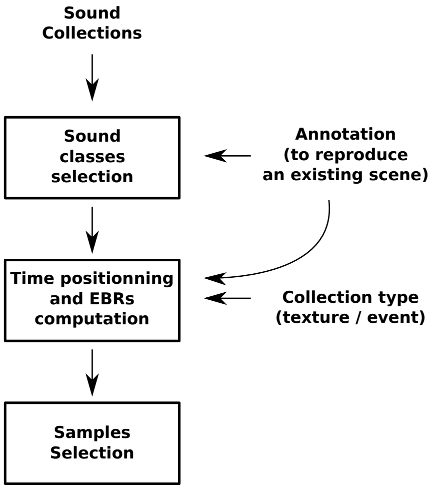
Sound Collections
Sound classes selection
Annotation (to reproduce an existing scene)
Time positionning and EBRs computation
Collection type (texture / event)
Samples Selection

Fig. 1. Schematic of the simulation process.

sound source. For the purpose of this study, the notion of a sound collection greatly overlaps the notion of a sound class, as this term is understood when tackling automatic detection tasks.

The resulting simulation model is depicted on Figure III-B. First, a number of sound sources or classes are selected, each of which is related to a specific sound collection. Second, simulation parameters are set, depending on the nature of the track (event or texture). Those parameters can also be estimated from pre-existing annotated recorded sound scenes. According to those parameters, the simulation process computes the number of samples used in each semantic track. Lastly, samples are randomly selected from the corresponding sound collection using a discrete uniform distribution.

### C. Model Formulation

The proposed model is source-driven as it uses sound tracks as basis elements. Each track gathers sounds coming from the same collection of either sound events or sound textures (see Figure III-B). After selecting the classes to be used, the putative recorded samples are sequenced to generate the sound environment. The sequencing process depends on the type of collection. Ideally, the sound collection design has to fulfill some perceptual constraints for the simulation to be ecologically valid, *e.g.* to produce realistic scenes, as described in Section II.

Considering that $s(n)$ is a given sound scene composed of $C$ sound classes $c_i$, the proposed model is such that:

$$s(n) = \sum_{i=1}^{C} t_i(n) \tag{1}$$

Where $n$ is the time index and $t_i$ is a sound track. For the sake of simplicity, we only detail here the model of an event track, then explain the adaptation of the model to a texture track.

A track $t_i$ is defined as a sequence of $n_i$ sound events $e_i^k(n)$ randomly chosen among the $|c_i|$ samples in class $c_i$ that is composed of recorded sound events $c_{i,m}$ with $1 < m < |c_i|$: for each $k$ in $[1..n_i]$, $e_i^k = c_{i,\mathcal{U}(1,|c_i|)}$, where $\mathcal{U}(x,y)$ represents an uniformly distributed integer random value between $x$ and $y$ included. Each event is normalized by its maximal amplitude and scaled by an amplitude factor sampled from a real normal distribution with average $\mu_i^a$ and variance $\sigma_i^a$, where the superscript $^a$ denotes an amplitude parameter. The interval separating the onset times of consecutive samples for track $i$ is, similarly, randomly chosen following a normal distribution with average $\mu_i^t$ and variance $\sigma_i^t$, where the superscript $^t$ denotes a timing parameter. Formally, each sequence $t_i$ is thus expressed as:

$$t_i(n) = \sum_{j=1}^{n_i} \mathcal{N}(\mu_i^a, \sigma_i^a) c_{i,\mathcal{U}(1,|c_i|)}(n - n_i^j) \tag{2}$$

$$n_i^j = n_i^{j-1} + \mathcal{N}(\mu_i^t, \sigma_i^t) \tag{3}$$

where $n_i^0$ is set to 0 by convention. The signal of an event is defined in such a way that $e(n) = 0$ if $n < 0$ or beyond the signal's duration.

In the case of a texture track, two implementation differences must be observed in order to maintain a perceptually coherent output: first, the signal amplitude is only drawn at random once, and that value is applied to all samples; second, sample start times are not randomized but chosen so that the texture recordings chosen from class $c_i$ will be played back-to-back with sufficient overlap to create an equal-power cross-fade between them, thus generating a continuous, seamless track.

## IV. Corpus simulation

This section describes the different corpora of simulated sound scenes built using the above described model for consideration in the experiments described in Section V. All the scenes are simulated using the annotations of the DCASE challenge test set of the “ Office Live ” (OL) task [16]. This dataset consists of scripted sequences of non-overlapping office sounds, originally recorded at Queen Mary University of London.

The root corpus is the test set considered in the DCASE challenge. It is called “ test-QMUL ” (testQ). This corpus is composed of 11 recordings of office scenes, each roughly one minute long. Scenes have been recorded in 5 different acoustic environments. The audio events present in the recordings have been divided into 16 sound event classes to be annotated: door knock, door slam, speech, laughter, clearing throat, coughing, drawer, printer, keyboard click, mouse click, object (specifically pen, pencil or marker) put on table surfaces, switch, keys (put on table), phone ringing, alert (beep sound) and page turn. Two different annotations coming from two distinct individuals have ben used as ground truth, thus leaving us with 22 scene-annotator couples. There is no time overlap between events.

Four corpora of simulated scenes are generated as depicted in Figure 2. They are respectively called “ instance-QMUL ” (insQ), “ abstract-QMUL ” (absQ), “ instance-IRCCYN ” (insI) and “ abstract-IRCCYN ” (absI). The labels “ QMUL ” and “ IRCCYN ” refer to two different datasets of event recordings used to generate the corpora, which have been recorded

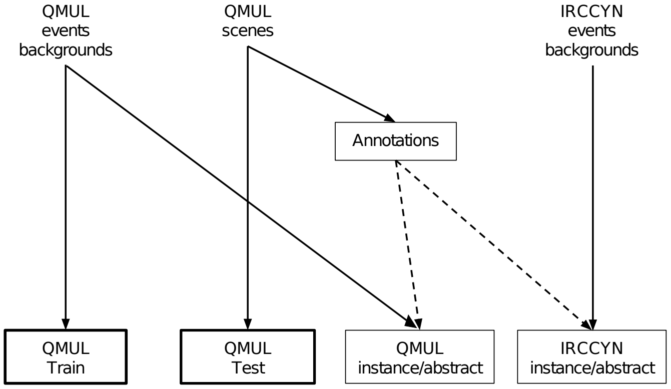

Fig. 2. Generation process of the corpora considered in this evaluation. As part of the DCASE challenge, systems were trained on QMUL Train and tested on QMUL Test during the DCASE challenge.

in different offices, at Queen Mary University of London (QMUL) and the Institute of Research on Communications and Cybernetics of Nantes (IRCCYN), respectively. The labels “instance” and “abstract” correspond to two distinct simulation processes, detailed below.

To generate the two QMUL corpora, we use recordings of audio events that have been extracted from recordings of isolated sounds done during the preparation of the DCASE challenge, but unused during the challenge, see [14] for further information on recording conditions. The extracted samples were therefore recorded in the same conditions than the test-QMUL corpus, but are not present in it. Depending on the considered sound class, 3 to 23 events per class are extracted. We also use event-free background recordings (texture) coming from the same acoustic environments than those of the test-QMUL corpus. These background recordings are used to generate the background noise (texture) of the instance-QMUL and abstract-QMUL corpora.

The two IRCCYN corpora are generated using recordings of isolated sound events with respect to the sound classes of the DCASE challenge. All recordings were performed at IRCCYN in a calm environment using the shotgun microphone AT8035 connected to a ZOOM H4n recorder. 20 samples of each class are used to generate the instance-IRCCYN and abstract-IRCCYN corpora, which corresponds to the cardinality of the DCASE train set in terms of event classes [14].

### A. “Instance” simulation process

The instance simulation process produces sound scenes with the same temporal structure and Event to Background power Ratios (EBRs) found in the corresponding scenes of the test-QMUL corpus. The EBR of an event of length $N$ (in samples) is obtained by computing the ratio in decibel between the event $E_{rms}$ and the background $B_{rms}$ root mean square measures:

$$EBR = 20log_{10}\left(\frac{E_{rms}}{B_{rms}}\right) \tag{4}$$

with

$$X_{rms} = \left(\frac{1}{N}\sum_{n=1}^{N} x(n)^2\right)^{1/2}$$

$x(n)$ may be replaced by $e(n)$ and $b(n)$, the sound pressures at sample $n$ of respectively the sound event and the background noise.

For each event of each scene-annotator couple of the test-QMUL corpus, the onset-offset times and an approximation of the EBR are considered. As it is not possible to isolate the background behind the events, the background level needed to compute the EBR is obtained using an event-free sequence of each real scene. These onsets-offsets and EBR are then used to generate the simulated scenes. For each simulated scene, at each onset of the corresponding annotator-couple scene, we randomly place an audio event belonging to the same audio class. To ensure that samples of recorded audio events are not too long compared to the annotated ones, recordings are cut off to the annotation length if the recording duration is larger than the annotation duration of at least 0.5 seconds.

Each event has its amplitude scaled to the same EBR than the test-QMUL corpus. Thus, the instance simulation process provides us with simulated scenes with temporal structures and sound levels that are close as possible as those of the real corpus test-QMUL.

### B. “Abstract” simulation process

For the abstract simulation process, the goal is to abstract the temporal structures and EBRs of the real scenes. To do so, the model described in Section III-C is instantiated using estimations of the $\mu_i^a$, $\sigma_i^a$, $\mu_i^t$ and $\sigma_i^t$ parameters; see eqs. (2) and (3). Estimation is done for each annotator-scene couple, using both the sound signals and the annotations of the test-QMUL corpus. To generate the simulated scenes, EBRs and time intervals between events are respectively obtained from the Normal distributions $\mathcal{N}(\mu_i^a, \sigma_i^a)$ and $\mathcal{N}(\mu_i^t, \sigma_i^t)$.

Similarly to the instance simulation process, event recordings are chosen randomly. For practical considerations, the start and termination times of the class sequence (sound track) are the same as the ones of the test-QMUL corpus. To ensure that the recorded samples are not significantly longer compared to the annotation times, the sample duration of a considered sound class $i$ has its duration $D$ thresholded as follows: $D - \mu_i^d - \sigma_i^d > 5$, with $\mu_i^d$ and $\sigma_i^d$ being respectively the average and standard deviation of the duration of the events belonging to the class $i$ in a given annotation, where the superscript $^d$ denotes a duration parameter. Setting the lower bound to 5 seconds allows us to minimize the impact of such operation on short impulsive sounds.

### C. Generated datasets

Five corpora called respectively test-QMUL (testQ - described in Section IV), instance-QMUL (insQ), abstract-QMUL (absQ), instance-IRCCYN (insI) and abstract-

IRCCYN (absI) are used for the evaluation described in Section V.

To measure the impact of different EBRs on the algorithm performance, the instance-QMUL corpus is composed of 4 sub-corpora called respectively "insQ-EBR_6", "insQ-EBR_0", "insQ-EBR_-6" and "insQ-EBR_-12". For the insQ-EBR_0 sub-corpus, the EBRs of test-QMUL scenes are preserved; for insQ-EBR_6, insQ-EBR_-6 and insQ-EBR_-12, the scenes are synthesized after adding offsets of $+6$dB, $-6$dB and $-12$dB, respectively, to the EBRs of the test-QMUL scenes.

For all the sub-corpora of the instance-QMUL corpus (insQ-EBR_6, insQ-EBR_0, insQ-EBR_-6 and insQ-EBR_-12) as well as the other corpora (abstract-QMUL, instance-IRCCYN, abstract-IRCCYN), each scene is simulated 10 times, each time with a different random selection of event and texture samples. Each of these corpora / sub-corpora is thus composed of 220 simulated scenes ($22 * 10$) corresponding to the 22 scene-annotator couples of the test-Q scenes replicated 10 times. All the datasets used are available online[1].

## V. Experiments

### A. User Study

In order to evaluate the acoustic realism of the simulated scenes, a user study is conducted with 15 participants that are asked to judge the acoustic realism from 1 (not realistic) to 7 (very realistic) of 22 scenes, half of them being recordings (QMUL-test) and half of them being simulated (IRCCYN-instance). The scenes are presented through headphones at a comfortable level. Subjects have to listen to the entire scene at least once before rating. Average ratings for the recorded scenes and simulated ones are respectively 4.4 and 3.3. From subject comments, it appears that the recorded scenes are not rated as very realistic because event occurrences are scripted and sometimes not well acted. Concerning the simulated ones, subjects reported that 1) the background is perceived as synthetic even though it was actually recorded in a quiet environment and 2) some events are trimmed. The latter issue is due to a design choice discussed in Section IV-A, taken to minimize the discrepancy between the simulated scene and the reference one. It should be noted that for many participants, some synthetic scenes were given a higher realism rating than some of the natural ones, which shows that while some noticeable differences can be made, they do not influence acoustic realism by a large margin.

### B. Evaluation Metric

The performance of event detection systems can be evaluated following several metrics; four are considered in the DCASE Challenge [9], [14], namely Acoustic Event Error Rate (AEER) [56], Precision, Recall, and F-measure. In order to improve the legibility of the following, we shall only retain F-measure, which proved during the initial challenge to be the most common and interpretable one.

Another variation is that those metrics can be computed over each frame or on event boundaries. In the latter case, the detection of the onset boundary can be considered solely or together with the offset. As annotating and consequently detecting the duration and the offset of events is notoriously difficult, we focus on the detection of the onset as the main objective, using a 100ms onset tolerance [9]. Furthermore, in order to achieve more comparable results across datasets and to ensure that repetitive events do not dominate the accuracy of an algorithm, the metric shall be class normalized. That is:

$$f = \frac{1}{C}\sum_{i=1}^{C} f_i \tag{5}$$

where $f_i$ is the F-measure achieved by the system while detecting event $i$. Thus, by considering the Class-Wise Event onset based F-measure (CWEBF) [9], performance evaluation is more invariant to event duration and distribution. We thus select this metric that was also collectively agreed upon by DCASE participants through the challenge requirements gathering stage.

### C. Detection systems

Together with a Baseline system provided by the organizers, 8 SED systems have been evaluated during the DCASE challenge. In the remaining, we follow the acronyms used in [9]. Those systems roughly follow the folowing processing chain: 1) pre-processing, 2)features extraction, 3) classification, and 4) post-processing, with some variety in the implementation of the different nodes. Features are most commonly Mel-Frequency Cepstral Coefficients (MFCCs) [57] but other sets of spectral features are also considered, with or without pre-processing such as denoising. The classifier of choice is the 2 layer Hidden Markov Model (HMM) [58] where the second layer models the transition between events. Other classifiers are also considered such as Random Forests (RF), Support Vector Machines (SVM) or Non-negative Matrix Factorization (NMF). Those algorithms have been trained on isolated events provided by the organizers of the challenge and tuned using a development set of annotated sound scenes.

All those algorithmic differences as well as their specific tuning result in specific behaviors that are important to evaluate in different testing conditions, especially those which evaluate their generalization capabilities.

### D. Results on QMUL datasets

With the kind permission of their authors, the above described systems are run on the simulated datasets on the same computing servers as the ones used for the challenge. Also, a rerun of the systems over the QMUL Test set has been done in order to ensure replication of the previously published challenge results [9].

Table I shows the class-wise event-based F-measure in percent achieved by the evaluated systems over the QMUL

[1] Dataset URLs:

- test-QMUL: https://archive.org/details/dcase2013_event_detection_testset_OL
- instance-QMUL, abstract-QMUL: https://archive.org/details/dcase_replicate_qmul
- instance-IRCCYN, abstract-IRCCYN: https://archive.org/details/dcase_replicate_irccyn

TABLE I
RESULTS OF THE EVALUATED SYSTEMS ON THE THREE QMUL DATASETS, IN TERMS OF CLASS-WISE EVENT-BASED F-MEASURE. RESULTS IN BOLD ARE EQUIVALENT PER ROW (PAIRED T-TEST AT 0.05 SIGNIFICANCE LEVEL) TO THE BEST PERFORMANCE PER ROW (DEPICTED WITH A *).

| system/dataset | testQ | insQ | absQ |
|---|---|---|---|
| Baseline | **9.0±4.8** | **10.5±3.0*** | **9.9±3.5** |
| CPS | **0.7±0.8** | **0.8±1.3** | **0.8±1.4*** |
| DHV | **30.7±8.4** | **34.5±7.5*** | **34.0±7.9** |
| GVV | **13.2±8.0** | **15.0±6.4*** | **14.6±6.2** |
| NR | **21.5±6.5*** | 6.8±5.7 | 7.4±5.8 |
| NVM | **28.2±5.9*** | 9.7±9.6 | 10.8±9.9 |
| NVM | **28.2±5.9*** | 9.7±9.6 | 10.8±9.9 |
| SCS | **41.5±7.6*** | **39.3±8.2** | **39.4±8.2** |
| VVK | **24.6±6.8*** | **19.7±8.7** | **19.2±9.2** |

TABLE II
MAXIMUM NUMBER OF FALSE POSITIVES FOR EACH SYSTEM, FOR THE THREE QMUL DATASETS (RESULTS ARE AVERAGED ACROSS RECORDINGS). THE CORRESPONDING EVENT CLASS IS DISPLAYED IN BRACKETS.

| System | testQ | insQ | absQ |
|---|---|---|---|
| Baseline | 3.14 (drawer) | 8.63 (drawer) | 7.40 (drawer) |
| CPS | 2.66 (door knock) | 9.04 (door slam) | 7.84 (door slam) |
| DHV | 8.44 (drawer) | 6.88 (drawer) | 8.01 (keyboard) |
| GVV | 3.08 (page turn) | 3.78 (page turn) | 3.55 (page turn) |
| NR | 4.33 (keyboard) | **25.35** (door slam) | **20.68** (door slam) |
| NVM | 1.26 (laughter) | **22.48** (cough) | **19.22** (cough) |
| SCS | 1.18 (alert) | 2.70 (drawer) | 1.72 (door slam) |
| VVK | 1.81 (alert) | 8.73 (door slam) | 8.20 (door slam) |

Test set and the simulated sets QMUL Instance and QMUL Abstract. The baseline, CPS, GVV and SCS systems performed equivalently across the 2 datasets. The DHV system performed better, but not by a significant margin. The VVK, NVM, and NR systems have their performance decreased; this decrease is significant for the NVM and NR systems (in terms of a paired t-test at 0.05 significance level), but not for the VVK one. The CPS system submitted to the DCASE Challenge had an implementation issue that prevented it to run correctly at the time of the challenge, giving poor results that are consistently replicated over the simulated datasets. For this reason, the CPS system will not be discussed further in the remaining of the paper. Leaving aside the NR and NVM systems, the ranking of the systems are equal for the 3 datasets. This result shows the usefulness of the proposed simulation scheme, since it is able to replicate and extend evaluation results achieved on the recorded dataset.

We now investigate further potential reasons explaining the behavior of the NVM and NR systems. In test mode, both systems first compute features and then run a classifier on them. Therefore, features were first checked for inconsistent values. The minimum and maximum values did not change across datasets, and the distribution of the features are indeed different across datasets, but not by a large margin.

Close inspection of the inter-class confusion matrices for each systems reveals that for the two systems the classification node may be responsible for this degradation of performance. Indeed, one event is triggered almost all the time which drastically increases the false alarm rate. This explains for a large part the decrease in performance of the NR and NVM systems. This behavior can easily be seen on Table V-D which displays the maximum number of false positives averaged over scenes of each datasets and their corresponding event; on the simulated datasets, the door slam event for NR and cough event for NVM are falsely triggered very often.

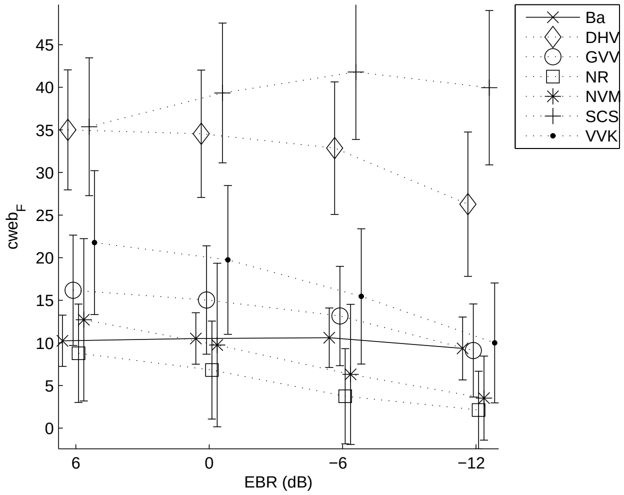

Fig. 3. Class wise event based F-measure (in percent) achieved by the systems on the QMUL instance datasets with varying EBR.

We conclude that this decrease in performance is most probably not due to some potential synthesis inconsistencies produced by the simulation process, but more due to an over fit of the classification node. Considering that both systems are the only submissions based on discriminative approaches, SVMs and RFs for the NR and NVM systems respectively, we may conjecture that the training framework of the DCASE challenge is not well suited for such classification schemes.

System performance following variations of the EBR is presented on Figure 3, where 0 dB of EBR roughly corresponds to the EBR level of the QMUL Test set. As expected, most systems see their performance decreasing with respect the decrease of the EBR. The ranking is preserved, and the spread between the 3 lowest performing systems greatly reduces at low EBR. The only system that does not follow this trend is the SCS system, which maintains a stable performance across all EBR ranges. This may be due to an effective signal enhancement which is an important pre processing node of this system [59].

### *E. Results on IRCCYN datasets*

When tackling a classification task, an important issue is whether the classification system under evaluation is able to generalize to unseen data whose annotation is consistent with the one used for training and tuning. To evaluate this generalization capability, it is useful to consider results achieved by the systems on the IRCCYN datasets, where the background and events are recorded in a different environment than the one used for recording the training data.

Whereas the expected behavior while considering the QMUL instance and abstract datasets was of equivalent performance compared to the ones achieved on test QMUL, the expected behavior with the IRCCYN dataset is a drop in performance, as can be seen on Figure 4, and this drop is significant for all the systems. More importantly, all the systems except the SCS one achieve similar performance when compared to the Baseline on the IRCCYN datasets, meaning that for most systems, the performance gain over the Baseline on the QMUL dataset may thus solely be due to an over adaptation of the system to the training data.

## VI. Discussion

To summarize the results discussed above, the use of the proposed model allows us to:

1) Replicate the ranking of systems in the same recording conditions for 5 systems among 7. The two problematic systems have their performance degraded most probably because of an overfit of the discriminative classifier they used.
2) Evaluate the generalization properties of the systems in new recording conditions. In this respect, the SCS system is the only one to generalize correctly.
3) Evaluate the robustness of the systems while facing different levels of background noise. Notably, once again, the SCS system exhibits good and stable performance across the EBR range, most probably because of an effective noise removal pre-processing step.

In light of those results, we believe that considering carefully designed simulated data is useful for gaining knowledge about the properties and behaviors of the systems under evaluation, thus helping designers in their algorithmic choices and their evaluation. Important factors influencing the performance such as the noise level, the level of polyphony, the intra-class diversity (acoustical difference between training and testing data) can be evaluated independently, without the burden of experimentally recording data with the desired properties and manually annotating them.

Even though the sole use of synthetic data for validating a computational approach is clearly not sufficient, we believe that the sole use of real data may not be sufficient either, should one wish to gain knowledge about the impact of some design and parametrization issues involved in the implementation of an engineering system. Indeed, real data is most of the time a scarce resource as the careful design of a large evaluation dataset is a very demanding task. Moreover, an *a posteriori* annotation of the presence of the events has to be performed by several humans whose agreement is not perfect and has to be mitigated.

We thus believe that considering simulated data is an in between approach, that together with final validation using real data is useful to get a better understanding about the systems under evaluation. The simulated sound scene datasets have been generated using a dedicated set of Matlab functions, which are publicly available[2].

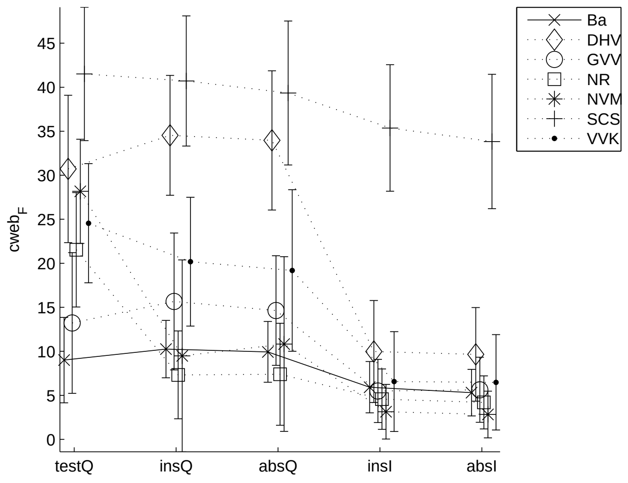

Fig. 4. Class-wise event based F-measure achieved by the different systems on the QMUL and IRCCYN datasets.

## VII. Conclusion

A morphological model of sound scenes has been presented. Following a collection-based approach, it generates a set of sound tracks which are sequences of event realizations drawn from specifically tailored sound sample collections. Its potential for generating simulated corpora of office event scenes is evaluated, by building upon the results obtained thanks to the IEEE DCASE challenge on the detection of sound events in an office environment.

Experiments carried out in this paper first demonstrate the capability of the proposed simulation scheme to generate data that is consistent with recorded datasets (QMUL corpora). Secondly, they show the capability of the simulation system to change only one aspect of the data at a time in order to evaluate 1) the robustness of the algorithms to higher levels of background noise and 2) the generalization capabilities of sound event detection systems when facing events and backgrounds recorded in a different acoustic environment (IRCCYN corpora).

We believe that considering those simulated corpora allows us to gain important knowledge about the behavior of the systems under evaluation. As most of the systems under evaluation were built for monophonic inputs (one event occurring at a given time), this paper focuses on modifying the acoustical properties of the background or the events. Future research will focus on the influence of the degree of overlap when facing polyphonic scenes, potentially with temporal interactions between events, both for single events (e.g. repetitions for a single event) and for interactions between event classes, as well as for the influence of room acoustics.

## Acknowledgment

The authors would like to thank Mark Plumbley for his support. We would also like to thank the participants of the DCASE OL and OS tasks, who allowed us to test their submitted systems for this paper.

[2] https://bitbucket.org/mlagrange/simscene/downloads

## References

[1] Ecology and acoustics: emergent properties from community to landscape. page 94, Paris, France, June 2014. Sueur, J. and Farina, A. and Bobryk, C. and Llusia, D. and McWilliam, J. and Pieretti, N., Musum national d'Histoire naturelle.

[2] Bryan C. Pijanowski, Almo Farina, StuartH. Gage, SarahL. Dumyahn, and Bernie L. Krause. What is soundscape ecology? an introduction and overview of an emerging new science. *Landscape Ecology*, 26(9):1213–1232, 2011.

[3] Steven R. Ness, Helena Symonds, Paul Spong, and George Tzanetakis. The orchive : Data mining a massive bioacoustic archive. *International Workshop on Machine Learning for Bioacoustics*, 2013.

[4] Dan Stowell and Mark D. Plumbley. Segregating event streams and noise with a markov renewal process model. *Journal of Machine Learning Research*, 14:2213–2238, 2013.

[5] Dan Stowell and Mark D. Plumbley. Large-scale analysis of frequency modulation in birdsong databases. *Methods in Ecology and Evolution*, 11, 2013.

[6] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen. Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1), 2013.

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

[8] Johnathan Turner Tae Hong Park, Michael Musick, Jun Hee Lee, Christopher Jacoby, Charlie Mydlarz, and Justin Salamon. Sensing urban soundscapes. In *EDBT/ICDT Workshops*, pages 375–382, 2014.

[9] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, October 2015.

[10] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[11] Meinard Müller. *Information Retrieval for Music and Motion*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[12] T Heittola, A Mesaros, A Eronen, and T Virtanen. Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013.

[13] C. Jacoby J. Salamon and J. P. Bello. A dataset and taxonomy for urban sound research. In *in Proc. 22nd ACM International Conference on Multimedia*, pages 158–161, Nov 2014.

[14] 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.

[15] E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley. Detection of overlapping acoustic events using a temporally-constrained probabilistic model. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 6450–6454, Shanghai, China, March 2016.

[16] Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events: An ieee aasp challenge. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2013.

[17] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo. The clear 2007 evaluation. In R. Stiefelhagen, R. Bowers, and J. Fiscus, editors, *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, pages 3–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[18] J.Vincent Barker, E., Ma, N., Christensen, C., Green, and P. The pascal chime speech separation and recognition challenge. *Computer Speech and Language*, 27(3), 2013.

[19] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni. The second 'chime' speech separation and recognition challenge: Datasets and tasks and baselines. In *ICASSP*, 2013.

[20] J. Barker, R. Marxer, E. Vincent, and S. Watanabe. The third 'chime' speech separation and recognition challenge: Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 504–511, Dec 2015.

[21] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos. The DIRHA simulated corpus. In *9th International Conference on Language Resources and Evaluation (LREC)*, pages 2629–2634, Reykjavik, Iceland, May 2014.

[22] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 7, October 2016.

[23] Israel Nelken and Alain de Cheveigné. An ear for statistics. *Nature neuroscience*, 16(4):381382, 2013.

[24] Sascha Spors, Heinz Teutsch, Achim Kuntz, and Rudolf Rabenstein. Sound field synthesis. In Yiteng Huang and Jacob Benesty, editors, *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, pages 323–344. Springer US, 2004.

[25] D.N. Zotkin, R. Duraiswami, and L.S. Davis. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia*, 6(4):553–564, Aug 2004.

[26] Charles Verron, Mitsuko Aramaki, Richard Kronland-Martinet, and Grégory Pallone. A 3-d immersive synthesizer for environmental sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1550–1561, 2010.

[27] Diemo Schwarz. State of the art in sound texture synthesis. In *Proc. Digital Audio Effects (DAFx)*, pages 221–231, 2011.

[28] Josh H. McDermott and Eero P. Simoncelli. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 71(5):926–940, September 2011.

[29] Richard Turner and Maneesh Sahani. Modeling natural sounds with modulation cascade processes. In *Advances in neural information processing systems*, pages 1545–1552, 2008.

[30] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[31] Joel S Snyder and Claude Alain. Toward a neurophysiological theory of auditory stream segregation. *Psychological bulletin*, 133(5):780, 2007.

[32] Valter Ciocca. The auditory organization of complex sounds. *Frontiers in bioscience: a journal and virtual library*, 13:148–169, 2007.

[33] Robert P Carlyon. How the brain separates sounds. *Trends in cognitive sciences*, 8(10):465–471, 2004.

[34] James A Ballas and James H Howard. Interpreting the language of environmental sounds. *Environment and behavior*, 19(1):91–114, 1987.

[35] Israel Nelken and Omer Bar-Yosef. Neurons and objects: the case of auditory cortex. *Frontiers in neuroscience*, 2(1):107, 2008.

[36] Danièle Dubois, Catherine Guastavino, and Manon Raimbault. A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories. *Acta Acustica united with Acustica*, 92(6):865–874, 2006.

[37] Manon Raimbault and Danile Dubois. Urban soundscapes: Experiences and knowledge. *Cities*, 22(5):339–350, October 2005.

[38] Maria Niessen, Caroline Cance, and Danile Dubois. Categories for soundscape: toward a hybrid classification. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2010, page 58165829, 2010.

[39] Catherine Guastavino. The ideal urban soundscape: Investigatng the sound quality of french cities. *Acta Acustica United with Acustica*, 92:945–951, 2006.

[40] Brian Gygi and Valeriy Shafiro. The incongruency advantage for environmental sounds presented in natural auditory scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):551, 2011.

[41] Maria E Niessen, Leendert Van Maanen, and Tjeerd C Andringa. Disambiguating sound through context. *International Journal of Semantic Computing*, 2(03):327–341, 2008.

[42] Olivier Houix, Guillaume Lemaitre, Nicolas Misdariis, Patrick Susini, and Isabel Urdapilleta. A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied*, 18(1):52–80, 2012.

[43] F. Guyot, M. Castellengo, and B. Fabre. *Catégorisation et Cognition: De la Perception au Discours*, chapter A study of the categorization of an everyday sound set, pages 41–58. Édition Kimé, Paris, France, 1997.

[44] Brian Gygi, Gary R Kidd, and Charles S Watson. Similarity and categorization of environmental sounds. *Perception & psychophysics*, 69(6):839–855, 2007.

[45] Michael M Marcell, Diane Borella, Michael Greene, Elizabeth Kerr, and Summer Rogers. Confrontation naming of environmental sounds. *Journal of clinical and experimental neuropsychology*, 22(6):830–864, 2000.

[46] N. J. Vanderveer. *Ecological acoustics: Human perception of environmental sounds*. Cornell University, Ithaca, NY, 1979.

[47] Eleanor Rosch and Barbara B Lloyd. Cognition and categorization. *Hillsdale, New Jersey*, pages 27–48, 1978.

[48] William W Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29, 1993.

[49] A.L. Brown, Jian Kang, and Truls Gjestland. Towards standardization in soundscape preference assessment. *Applied Acoustics*, 72(6):387–392, May 2011.

[50] Michael Southworth. The sonic environment of cities. *Environment and behavior*, 1969.

[51] Valérie Maffiolo. Semantic and acoustic characterization of urban environmental sound quality. *Ph. D. dissertation, Université du Maine, France*, 1999.

[52] Josh H McDermott, Michael Schemitsch, and Eero P Simoncelli. Summary statistics in auditory perception. *Nature neuroscience*, 16(4):493–498, 2013.

[53] Nicolas Saint-Arnaud. *Classification of sound textures*. PhD thesis, Massachusetts Institute of Technology, 1995.

[54] Nicolas Saint-Arnaud and Kris Popat. Analysis and synthesis of sound textures. In *in Readings in Computational Auditory Scene Analysis*. Citeseer, 1995.

[55] Trevor R. Agus, Simon J. Thorpe, and Daniel Pressnitzer. Rapid formation of robust auditory memories: Insights from noise. *Neuron*, 66(4):610–618, May 2010.

[56] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan. The clear 2006 evaluation. In Rainer Stiefelhagen and John Garofolo, editors, *Multimodal Technologies for Perception of Humans*, volume 4122 of *Lecture Notes in Computer Science*, pages 1–44. Springer Berlin Heidelberg, 2007.

[57] Steven Davis and Paul Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[58] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[59] J. Schröder, B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze. Acoustic event detection using signal enhancement and spectro-temporal feature extraction. Technical report, 2013. http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/abstracts/OL/SCS.pdf.

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

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

**Mathias Rossignol** obtained his PhD from Rennes University (France) in 2005, with a specialty in Natural Language Processing. He then focused his research on speech analysis and recognition at the International Research Center MICA (Hanoi, Vietnam), before moving on to Auditory Scene Analysis at IRCAM (Institute for Acoustic / Music Research and Coordination, Paris, France).

**Emmanouil Benetos** (S'09-M'12) received the B.Sc. and M.Sc. degrees in informatics from the Aristotle University of Thessaloniki, Greece, in 2005 and 2007, respectively, and the Ph.D. degree in electronic engineering from Queen Mary University of London, U.K., in 2012. From 2013 to 2015, he was a University Research Fellow with the Department of Computer Science, City University London, London, U.K. He is currently a Royal Academy of Engineering Research Fellow with the Centre for Digital Music, Queen Mary University of London, U.K. His research focuses on signal processing and machine learning for music and audio analysis, as well as applications to music information retrieval, acoustic scene analysis, and computational musicology.

**Axel Roebel** received the Diploma in electrical engineering from Hannover University in 1990 and the Ph.D. degree (summa cum laude) in computer science from the Technical University of Berlin in 1993. In 1994 he joined the German National Research Center for Information Technology (GMD-First) in Berlin where he continued his research on adaptive modeling of time series of nonlinear dynamical systems. In 1996 he became assistant professor for digital signal processing in the communication science department of the Technical University of Berlin. In 2000 he obtained a research scholarship to pursue his work on adaptive sinusoidal modeling at CCRMA Standford University, and in the same year he joined IRCAM for working in the analysis-synthesis team doing research on frequency domain signal processing. In summer 2006 he was Edgar-Varse guest professor for computer music at the Electronic studio of the Technical University of Berlin. Since 2011 he is head of the analysis/synthesis team of IRCAM. His current research interests are related to music and speech signal modeling, transformation and synthesis.