Detection and Classification of Acoustic Scenes and Events

Giannoulis, D; Benetos, E; Stowell, D; Rossignol, M; Lagrange, M; Plumbley, MD
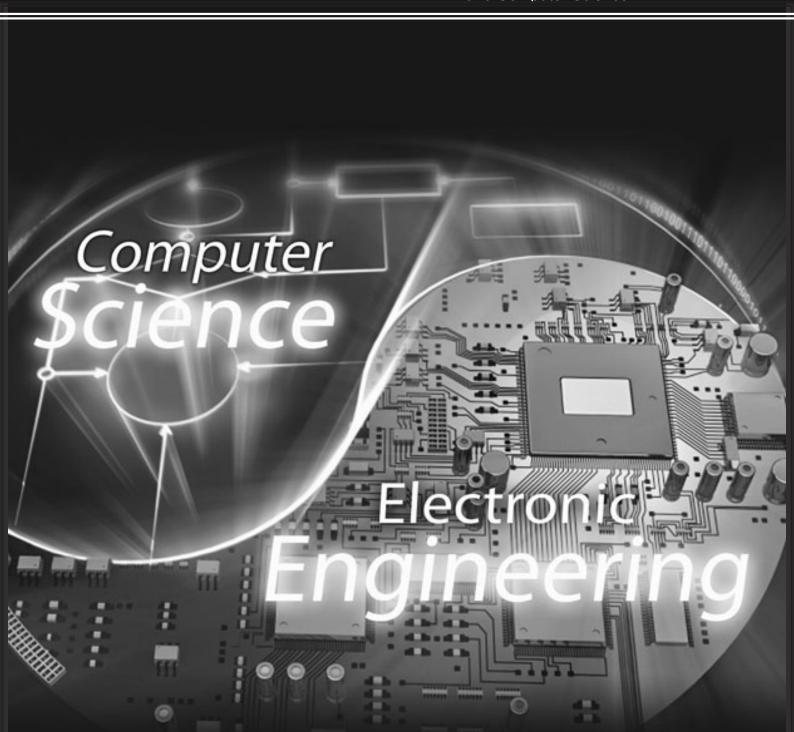
Detection and Classification of Acoustic Scenes and Events

Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange and Mark Plumbley

Queen Mary
University of London

An IEEE AASP Challenge

Detection and Classification of Acoustic Scenes and Events (Technical Report)

Dimitrios Giannoulis[†], Emmanouil Benetos[§], Dan Stowell[†], Mathias Rossignol[‡], Mathieu Lagrange[‡] and Mark Plumbley[†]

I. INTRODUCTION

Over the last decade, there has been an increased interest in the speech and audio processing community in code dissemination and public evaluation of proposed methods. Public evaluation can serve as a reference point for the performance of proposed methods and can also be used for studying performance improvements throughout the years. For example, source separation and automatic music transcription have been well-defined, they have their own performance metrics established, and public evaluations are performed for each - (see the SiSEC evaluation for signal separation [1] for the 1st and the MIREX competition for music information retrieval [2] for the 2nd). However, for researchers working on the field of computational auditory scene analysis (CASA) [3] and specifically, on the tasks of modeling and identifying acoustic scenes containing non-speech and non-music and detecting audio events, there is not yet a coordinated established international challenge in this area. We therefore propose to organise a challenge on the performance evaluation of systems for the detection and classification of acoustic events. This challenge will help the research community move a step forward in better defining the specific task and will also provide incentive for researchers to actively pursue research on this field. Finally, it will help shedding light on controversies that currently exist in the task and offer a reference point for systems developed to perform parts of this task.

We should mention that at present the closest challenge to the one we propose is TRECVID Multimedia Event Detection, where the focus is on audiovisual, multi-modal event detection in video recordings [4]. There are researchers that are using only the audio from the TRECVID challenge in order to evaluate their systems but a dataset explicitly developed for audio challenges would offer a much better evaluation framework since it would be much more varied with respect to audio. In addition, such a dataset would be made so that it would address the needs for a more thorough evaluation of audio analysis systems and would potentially be used more widely and set itself as a standard. We should also note that a public evaluation on Audio Segmentation and Speaker Diarization [5] has also been proposed. This proposed evaluation task consists of segmenting a broadcast news audio document into a few specific classes that are: music, speech, speech with music/noise in background or other. Therefore it is addressing a very specific task and it does not overlap with the proposed challenge.

Finally, one public evaluation that is related to the proposed challenge took place in 2006 and 2007, as part of the CLEAR evaluations [6], funded by the CHIL project. Several tasks on audio-only, video-only or multimodal tracking and event detection were proposed and among them was an evaluation on "Acoustic Event Detection and Classification". The datasets were recorded during several interactive seminars and contain events related to seminars (speech, applause, chair moving, etc). From the datasets created for the evaluations, the "FBK-Irst database of isolated meeting-room acoustic events" [7] has widely been used in the event detection literature; however, the

† The authors are with the Centre for Digital Music, Queen Mary University of London, Mile End Rd., London E1 4NS, UK. E-mail: {dimitrios, dans, markp}@eecs.qmul.ac.uk

§ The author is at the Department of Computer Science, City University London, Northampton Square, London EC1V 0HB, UK. E-mail: emmanouil.benetos.1@city.ac.uk

‡ The authors are with the Sound Analysis/Synthesis Team, IRCAM, 1 place Igor stravinsky, 75004, Paris, France. E-mail: mathieu.lagrange@ircam.fr, mathias.rossignol@gmail.com

aforementioned dataset contains only non-overlapping events. The CLEAR evaluations, although promising and innovative at the time, did not lead to the establishment of a widely-accepted evaluation challenge for this type of tasks mainly because the datasets were limited to specific types of events and acoustic scenes. These evaluations have been discontinued with the end of the CHIL project.

II. Background

Computational auditory scene analysis (CASA) includes a wide set of algorithms and "machine listening" systems that deal with the analysis of acoustic scenes. Most of them model to some extent the human auditory system and its mechanisms and aim to detect, identify, separate and segregate sounds in the same way that humans do [3]. Two closely related tasks in computational auditory scene analysis (CASA) are *acoustic scene classification* and *detection of sound events within a scene* [3]. A system involved in the first task has as a goal to characterise or "label" the environment in which the audio was recorded by providing a semantic label [8], whereas a system aiming to detect sound events is trying to segment the audio in pieces that represent a single occurrence of a specific event class by estimating the start and end time of each event and if necessary separating it from other overlapping events.

Acoustic scene classification aims to characterize the environment of an audio stream by providing a semantic label [8]. It can be conceived of as a standard classification task in machine learning: given a relatively short clip of audio, the task is to select the most appropriate of a set of scene labels. There are two main methodologies found in the literature. One is to use a set of low-level features under a bag-of-frames approach. This approach treats the scene as a single object and aims at representing it as the long-term statistical distribution of some set of local spectral features. Prevailing among different features for the approach is the Mel-frequency Cepstral Coefficients (MFCCs) that have been found to perform quite well [8].

The other is to use an intermediate representation prior to classification that models the scene using a set of higher level features that are usually captured by a vocabulary or dictionary of "acoustic atoms". These atoms usually represent acoustic events or streams within the scene which are not necessarily known a priori and therefore are learned in an unsupervised manner from the data. Sparsity or other constraints can be adopted to lead to more discriminative representations that subsequently ease the classification process. An example is the use of non-negative matrix factorization (NMF) to extract bases that are subsequently converted into MFCCs for compactness and used to classify a dataset of train station scenes [9]. Building upon this approach, the authors in [10] used shift-invariant probabilistic latent component analysis (SIPLCA) with temporal constrains via hidden Markov models (HMMs) that led to improvement in performance. In [11] a system is proposed that uses the matching pursuit algorithm to obtain an effective time-frequency feature selection that are afterwards used as supplement to MFCCs to perform environmental sound classification.

The goal of *acoustic event detection* is to label temporal regions, such that each represents a single event of a specific class. Early work in event detection treated the sound signal as monophonic, with only one event detectable at a time [12]. Events in a typical sound scene may co-occur, and so polyphonic event detection, with overlapping event regions, is desirable. However, salient events may occur relatively sparsely and there is value even in monophonic detection. There has been some work on extending systems to polyphonic detection [13]. Event detection is perhaps a more demanding task than scene classification, but at the same time heavily intertwined. For example, information from scene classification can provide supplementary contextual information for event detection [14]. Many proposed approaches can be found in the literature among which spectrogram factorization techniques tend to be a regular choice. In [15] a probabilistic latent semantic analysis (PLSA) system, a closely related approach to NMF, was proposed to detect overlapping sound events. In [16] a convolutive NMF algorithm applied on a Mel-frequency spectrum was tested on detecting non-overlapping sound events. Finally, a number of proposed systems focus on the detection and classification of specific sound events from environmental audio scenes such as speech [17], birdsong [18], musical instrument and other harmonic sounds [19] or pornographic sounds [20].

III. Challenge Description

The aim of the proposed challenge is to build a specific set of sub-challenges for the detection and classification of acoustic scenes and events in monaural recordings. Our goal is to provide a focus of attention for the scientific

community in developing systems for CASA that will encourage sharing of ideas and improve the state of the art, potentially leading to the development of systems that achieve a performance close to that of humans.

The first challenge addresses the problem of identifying and classifying acoustic scenes and soundscapes. The second challenge addresses the problem of identifying individual sound events that are prominent in an acoustic scene. Two distinct experiments take place for sound event identification, one for simple acoustic scenes without overlapping sounds and the other using complex scenes in a polyphonic scenario. In an everyday scenario, most of the sounds that reach our ears tend to stem from a multitude of sources so the polyphonic case would be more interesting but much more challenging.

IV. EVALUATION DATA

A. Datasets

There are three datasets overall, one for scene classification and two for event detection, which will be described below.

1) Scene Classification Dataset: The dataset for the scene classification (SC) challenge consists of 30 sec recordings of various acoustic scenes. The scene classification dataset consists of 2 equally proportioned parts each made up of 10 audio recordings for each scene (class), for a total of 100 recordings per part. One for public release (available to participants to build up and investigate the performance of their system), and one private set for evaluating submissions. The public dataset is published on the C4DM Research Data Repository (accessible through [21]). Scenes include:

- busy street
- quiet street
- supermarket/store
- restaurant
- office

- park
- bus
- tube/metro
- tubestation
- open market

For each scene type, three different recordists (DG, DS, EB) visited a wide variety of locations in Greater London over a period of months (Summer and Autumn 2012), and in each scene recorded a few minutes of audio. We ensured that no systematic variations in the recordings covaried with scene type: all recordings were made in moderate weather conditions, and varying times of day and week, and each recordist recorded each scene type. 30-second segments were selected after careful review of the recordings to ensure they were free of issues such as mobile phone interference or microphone handling noise.

2) Event Detection - Office Live Dataset: The dataset for the Event Detection - Office Live task consists of three subsets (a training, a development and a testing dataset). The training set[1] contains instantiations of individual events for every class. The development (validation) and testing dataset, denoted as office live (OL), consists of roughly 1 min long recordings of every-day audio events in a number of office environments (different-size and absorbing quality rooms, different number of people in the room and varied noise level). The audio events include:

- door knock
- door slam
- speech
- human 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
- short alert (beep) sound
- page turning

Since there is inherent ambiguity in the annotation process, there are available two different annotations corresponding to two different human annotators. Annotators were trained in Sonic Visualiser[2] to use a combination of

---

[1]The training set for the Event Detection OL and OS tasks can be obtained from http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/28

[2]http://www.sonicvisualiser.org/

listening and inspecting waveforms/spectrograms to refine the locations. We then examined the two annotations per recording for for consistency and any instances of errors. Especially in the case of long soft tails in the offset of some events it is humanly impossible to extract a meaningful and accurate offset point and it usually comes down to the subjective opinion of the annotator where the offset for that event is. Therefore, including more than one annotation helps generalise the evaluation of the systems by allowing a small trade-off in the complexity of the testing process. Participants are welcome to use both, the average or only one of the two. Evaluation is made using both annotations.

The training set includes 24 recordings of individual sounds per class, followed by annotations of their onset and offset in sec. The development set includes 3 recordings of a series of events from one of the office environments. These recordings are also accompanied by annotation of the events' onsets and offsets. The third set that is not released contains recordings of sound events made in all office environments, excluding the one used for the development set.

3) Event Detection - Office Synthetic Dataset: The third dataset (for the Event Detection - Office Synthetic task) contains artificially sequenced sounds provided by the Analysis-Synthesis team of IRCAM. The aim of this subtask is to study the behavior of tested algorithms when facing different levels of complexity such as the event to background energy ratio, the level of overlap between individual events etc. The benefit of using such a dataset is that the experiment is more controllable and practical than utilizing real recordings. In addition to that, the ground truth is most accurate even for polyphonic mixtures with lots of overlaps among different sounds. We expect systems to perform better in this dataset but it could help for measuring the performance of systems in artificially created recordings compared to real recordings.

The data for the OS task consist of three subsets like the previous task. The training dataset consists of audio recordings of individual events which are identical to the one for the realistic task. The development and testing datasets consist of artificial scenes built by sequencing recordings of individual events (different recordings from the ones used for the training dataset) and background recordings provided by C4DM. As the data are recorded by QMUL specifically for this challenge, confidentiality is ensured.

A scene synthesizer able to easily create a large set of acoustic scenes from many recorded instances of individual events was designed. The synthetic scenes are generated by randomly selecting for each occurrence of each event we wish to include in the scene one representative excerpt from the natural scenes, then mixing all those samples over a background noise. The distribution of events in the scene is also random, following high-level directives that specify the desired density of events. The average SNR of events over background noise is also specified and, unlike in the natural scenes, is the same for all event types (this is a deliberate decision). The synthesized scenes are mixed down to mono in order to avoid having spatialization inconsistencies between successive occurrences of a same event; spatialization including room reverberation is left for future work. The resulting development and testing datasets consist of scripted/synthetic sequences with varying durations, with accompanying ground-truth annotations. The development dataset is published on the C4DM Research Data Repository (accessible through [21]).

Note: All datasets for the challenge are released under a creative commons (CC BY) license[3].

B. Recording Equipment

The Centre for Digital Music at Queen Mary University of London collected data on environmental audio to be used exclusively for the challenge. The recording equipment includes two settings. The first is a high-quality Soundfield microphone system, model SPS422B [22], able to capture 4-channel surround sound with high clarity that can also be mapped to stereo or mono in a later state if necessary. The second is a set of Soundman binaural microphones, model OKM II [23], specifically made so that they imitate a pair of in-ear headphones that the user can wear. The portability and subtlety of that system enables the user not to attract any attention from people in the environment and therefore, we can obtain everyday recordings unobstructed and with relative ease. Furthermore, the recorded audio is very similar to the sound that reaches the human auditory system of the person wearing the equipment as it is recorded after being filtered by the head-related transfer function (HRTF) [24]. Therefore, the resulting data carry also binaural information about the sound that could additionally be utilized as cues for sound

---

[3]For more details on licensing please visit: http://creativecommons.org/licenses/

event and scene detection from audio or simply be ignored entirely by adding the two channels together in order to obtain a mono recording.

The sound files for the 1st task (scene classification), recorded with the binaural microphones, have the following specifications: PCM, 44100 Hz, 16 bit, two-channel (CD quality). The specifications for the sound files for the other tasks, that were recorded with the Soundfield microphone system, are: two-channel stereo (mixed down from 4-channel B-format), 44100 Hz, 24 bit. The B-format is also released together with the stereo versions but the challenge is run on stereo and not the B-format. The participants for the challenge have the flexibility to mix recordings down to mono if they desire to do so.

Finally it should be noted that the recording level was held constant for all sounds in both the training and test recordings and for all tasks.

V. METRICS

A. Scene Classification

For classifying acoustic scenes, the output of each run for a single file only contains the class label. As in the MIREX train/test tasks [2], the metrics that are computed are the raw classification (identification) accuracy, a normalized classification accuracy per class, the standard deviation, and a confusion matrix for each submission. For this train/test task, participating algorithms are evaluated using 5-fold cross validation.

B. Event Detection

For event detection, three types of evaluations take place. A frame-based, an event-based, and a class-wise event-based evaluation. We believe that both methods together can provide a thorough assessment of the various systems, with the event-based evaluation capturing the accuracy of the overall event detection, and the frame-based evaluation offering in finer detail the accuracy over time for each system.

The output of each run is a file that should contain the onset, offset and the event ID separated by a tab, ordered in terms of onset times:

onset   offset   E01
onset   offset   E02
onset   offset   E03
...

Frame-based evaluation is performed using a 10ms step. The main metric utilised for the frame-based evaluation is a frame-based version of the acoustic event error rate [7]:

$$AEER = (D + I + S)/N \cdot 100 \tag{1}$$

where N is the number of events to detect for that specific frame, D is the number of deletions (missing events), I is the number of insertions (extra events), and S is the number of event substitutions, defined as $S = \min\{D, I\}$. Frame-level metrics are averaged over time for the duration of the recording.

Additional metrics are given by using the Precision, Recall, and F-measure (P-R-F). By denoting as r, e, and c the number of ground truth, estimated and correct events for a given 10ms frame, the aforementioned metrics are defined as:

$$Pre = \frac{c}{e}, \quad Rec = \frac{c}{r}, \quad F = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec} \tag{2}$$

For the onset-only event-based evaluation, each event is considered to be correctly detected within a 100ms tolerance window. For the onset-offset event-based evaluation, each event is correctly detected if its onset is within a 100ms tolerance window and its offset is within 50% range of the ground truth event's offset with respect to the duration of the event. As in the frame-based task, the AEER and P-R-F metrics for both the onset-only and the onset-offset event detection tasks can be defined accordingly. It should also be noted that duplicate events are considered as false alarms.

Finally, a class-wise event-based evaluation also takes place, in order to ensure that that repetitive events do not dominate the accuracy of an algorithm. The output of the algorithm is the same as in the event-based evaluation,

but in this fase the AEER and P-R-F metrics are computed for each class separately within a recording and will be averaged accoss a recording. For example, the class-wise F-measure is defined as:

$$F = \sum_k F_k / K \tag{3}$$

where F_k denotes the computed F-measure taking into account detected events for class k.

VI. SCHEDULE

The schedule for the proposed challenge is as follows:

1) **June 2012:** An open call for participation and discussion for the challenge will be announced in related mailing lists (AUDITORY, IEEE SPS Newsletter, IEEE AASP members and affiliates, machine-listening) and will also be advertised in related conferences (e.g. MLSP). The challenge website will be updated accordingly.

2) **August 2012**: Deadline for encouraging participants to contribute in the discussions regarding the challenge specifications (a mailing list will be created for any challenge-related discussions).

3) **March 2013**: Deadline for code submission. The code can be either run by the challenge organisers or by the participants themselves. The code should be accompanied by a maximum 3-page description of their work, in the IEEE double-column conference format (templates will be uploaded in the challenge website, copyright will remain with the authors).

4) **May 2013**: Submission deadline for WASPAA 13. Authors of novel work related to the challenge are encouraged to submit regular papers to the workshop.

5) **October 2013**: The 3-page descriptions will be made public along with the evaluation results. Authors are invited to submit camera-ready versions of the descriptions, reflecting the results of the evaluation. During WASPAA 13, each submission will be presented by the participants during one of the regular poster sessions. A 20min oral presentation and discussion regarding the specific challenge will also take place in the same workshop[1].

6) **November 2013**: Invite select participants to submit novel work for the challenge in an IEEE TASLP/JSTSP special issue on the AASP challenges. The challenge organisers will also write an overview article on the challenge and the current trends in the field (this overview article can also be part of a Signal Processing Magazine submission, in order to increase visibility).

REFERENCES

[1] E. Vincent, S. Araki, F.J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B.V. Gowreesunker, D. Lutter, and N.Q.K. Duong, "The Signal Separation Evaluation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

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

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

[4] "TRECVID 2011 MED Evaluation track," http://www.nist.gov/itl/iad/mig/med11.cfm.

[5] "Albayzin 2010 audio segmentation and speaker diarization evaluation task," http://fala2010.uvigo.es/index.php?option=com_content&view=article&id=60\%3Aaass&catid=36&Itemid=65&lang=en.

[6] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," *Multimodal Technologies for Perception of Humans*, pp. 1–44, 2007.

[7] "FBK-Irst database of isolated meeting-room acoustic events," http://catalog.elra.info/product_info.php?products_id=1093, 2008, European Language Resources Association.

[8] 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*, vol. 122, pp. 881, 2007.

[9] B. Cauchi, "Non-negative matrix factorisation applied to auditory scenes classification," MS thesis, 2011.

[10] E. Benetos, M. Lagrange, and S. Dixon, "Characterization of acoustic scenes using a temporally-constrained shift-invariant model," in *Proc DAFX, York, UK*, 2012.

[11] S. Chu, S. Narayanan, and C.-C. Jay Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Trans Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[12] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc EUSIPCO*, 2010.

[13] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Proc CHiME*, 2011, pp. 36–40.

---

[1]After contacting the WASPAA 2013 chairs, it has been agreed to assign a time slot for presenting the challenge and its results. The exact date and form of the session is to be finalised together with the workshop schedule when this is announced.

[14] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, 2013.

[15] A. Mesaros, T. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection," in *Proc EUSIPCO*, 2011, pp. 1307–1311.

[16] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *Proc WASPAA*, 2011, pp. 69–72.

[17] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, no. 1, pp. 5–25, 2005.

[18] F. Briggs, B. Lakshminarayanan, et al., "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *Journal of the Acoustical Society of America*, vol. 131, pp. 4640–4650, 2012.

[19] D. Giannoulis, A. Klapuri, and M. D. Plumbley, "Recognition of harmonic sounds in polyphonic audio using a missing feature approach," in *Proc ICASSP (to appear)*, 2013.

[20] M. J. Kim and H. Kim, "Automatic extraction of pornographic contents using radon transform based audio features," in *CBMI*, 2011, pp. 205–210.

[21] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," an IEEE AASP Challenge, 2013, www.elec.qmul.ac.uk/digitalmusic/sceneseventschallenge/.

[22] "Soundfield SPS422B Microphone System," http://www.soundfield.com/products/sps422b.php.

[23] "SoundMan Binaural Microphone system," http://dev.soundman.de/.

[24] C.I. Cheng and G.H. Wakefield, "Introduction to head-related transfer functions (HRTFs): representations of HRTFs in time, frequency, and space," *Journal of the Audio Engineering Society*, vol. 49, no. 4, Apr. 2001.