A DATABASE AND CHALLENGE FOR ACOUSTIC SCENE CLASSIFICATION AND EVENT DETECTION

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

† Centre for Digital Music, Queen Mary University of London, London, UK.
‡ Department of Computer Science, City University London, London, UK.
§ Sound Analysis/Synthesis Team, IRCAM, Paris, France.

ABSTRACT

An increasing number of researchers work in computational auditory scene analysis (CASA). However, a set of tasks, each with a well-defined evaluation framework and commonly used datasets do not yet exist. Thus, it is difficult for results and algorithms to be compared fairly, which hinders research on the field. In this paper we will introduce a newly-launched public evaluation challenge dealing with two closely related tasks of the field: acoustic scene classification and event detection. We give an overview of the tasks involved; describe the processes of creating the dataset; and define the evaluation metrics. Finally, illustrations on results for both tasks using baseline methods applied on this dataset are presented, accompanied by open-source code.

Index Terms— Computational auditory scene analysis, acoustic scene classification, acoustic event detection

1. INTRODUCTION

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 [1].

Certain practical applications that fall under the umbrella of CASA, such as noise-robust automatic speech recognition and automatic music transcription, have seen a high amount of research over the last decades, and state-of-the-art approaches for both are able to achieve satisfactory performance, comparable to that of humans (see the MIREX evaluation[1] and the CHiME challenge[2]). However, the field of CASA involves a much wider set of tasks and "machine listening" systems, many of which are far from being fully explored at a research level yet.

Over the last few years the tasks of identifying auditory scenes, and that of attempting to detect and classify individual sound events within a scene, have seen a particular rise in research, mainly due to them being interdependent with other tasks of high interest such as blind source separation. Despite an increasing number of attempts by the community for code dissemination and public evaluation of proposed methods [2, 3, 4], it is evident that there is not yet a coordinated, established, international challenge in this particular area with a thorough set of evaluation metrics that fully define the two tasks. By organising the present "IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events" [5] we aim to do exactly that. In the rest of the paper, we present the datasets created, the evaluation metrics used, and provide evaluation results using two baseline methods.

2. BACKGROUND

Acoustic scene classification aims to characterize the environment of an audio stream by providing a semantic label [6]. 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 [6]. 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

This work has been partly supported by ESPRC Leadership Fellowship EP/G007144/1, by EPSRC Grant EP/H043101/1 for QMUL, and by ANR-11-JS03-005-01 for IRCAM. D.G. is funded by a Queen Mary University of London CDTA Research Studentship. E.B. is supported by a City University London Research Fellowship.

[1] http://music-ir.org/mirexwiki/

[2] http://spandh.dcs.shef.ac.uk/projects/chime/challenge.html

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 [7]. Building upon this approach, the authors in [8] used shift-invariant probabilistic latent component analysis (SIPLCA) with temporal constrains via hidden Markov models (HMMs) that led to improvement in performance. In [9] 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 [10]. 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 [11]. 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 [12]. Many proposed approaches can be found in the literature among which spectrogram factorization techniques tend to be a regular choice. In [13] a probabilistic latent semantic analysis (PLSA) system, a closely related approach to NMF, was proposed to detect overlapping sound events. In [14] 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 [15], birdsong [16], musical instrument and other harmonic sounds [17] or pornographic sounds [18].

3. CHALLENGE

This section presents the proposed IEEE-sponsored challenge in acoustic scene classification and event detection [5]. Firstly, the datasets for the two aforementioned tasks are described, followed by definitions on the employed evaluation metrics.

3.1. Scene classification datasets

In order to evaluate Scene Classification systems we created a dataset across a pre-selected list of scene types, representing an equal balance of indoor/outdoor scenes in the London area: *bus*, *busystreet*, *office*, *openairmarket*, *park*, *quietstreet*, *restaurant*, *supermarket*, *tube*, *tubestation*. To enable participants to further explore whether machine recognition could benefit from the stereo field information available to human listeners [1, Chapter 5], we recorded in binaural stereo format using a Soundman OKM II microphone.

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.

We then reviewed the recordings to select 30-second segments that were free of issues such as mobile phone interference or microphone handling noise, and collated these segments into two separate datasets: one for public release, and one private set for evaluating submissions. The segments are 30-second WAV files (16 bit, stereo, 44.1 kHz), with scene labels given in the filenames. Each dataset contains 10 examples each from 10 scene types. The public dataset is published on the C4DM Research Data Repository (accessible through [5]).

3.2. Event detection (office) datasets

For the Event Detection task, we addressed the problem of detecting acoustic events in an office enviromnent. In order to control the degree of polyphony in the dataset, so that algorithms' performance can be evaluated using different polyphony levels, we followed two related approaches: we recorded live, scripted, monophonic sequences in real office environments; and we also recorded isolated events as well as background ambience, and artificially composed these into scenes with controllable polyphony.

For the scripted recordings, we created scripts by random ordering of event types, and then recruited a variety of paid participants to perform the scripts in various office rooms within QMUL. For each script, multiple takes were used, and we selected the best take as the one having the least amount of unscripted background interference. Event types used were: *alert* (short alert (beep) sound), *clearthroat* (clearing throat), *cough*, *doorslam* (door slam), *drawer*, *keyboard* (keyboard clicks), *keys* (keys put on table), *knock* (door knock), *laughter*, *mouse* (mouse click), *pageturn*, (page turning), *pendrop* (pen, pencil, or marker touching table surfaces), *phone*, *printer*, *speech*, *switch*. To capture the spatial layout of the acoustic environment, recordings were made in first order B-format with a Soundfield model SPS422B microphone placed in an open space in the room, with events spatially distributed around the room. Recordings were mixed down to stereo (using the common "Blumlein pair" configuration). The challenge is conducted using the stereo files, with scope for future challenges to be extended to full B-format and take

into account spatial information for event detection.

Since there is inherent ambiguity in the annotation process, we recruited two human annotators to annotate the onset and offset times of events in the recordings. Annotators were trained in Sonic Visualiser[3] to use a combination of listening and inspecting waveforms/spectrograms to refine the locations. We then inspected the two annotations per recording for any large discrepancies, which allowed us to detect any instances of error. The remaining small deviations between the annotations reflect the ambiguity in event boundaries.

For the second approach, we designed a scene synthesizer able to easily create a large set of acoustic scenes from many recorded instances of individual events. 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 [5]).

3.3. Challenge evaluation metrics

For the scene classification task, participating algorithms will be evaluated with 5-fold stratified cross validation. The raw classification (identification) accuracy, standard deviation and a confusion matrix for each algorithm will be computed.

For the event detection tasks, in order to provide a thorough assessment of the various systems three types of evaluations will take place, namely a frame-based, event-based, and class-wise event-based evaluation. Frame-based evaluation is performed using a 10ms step and metrics are averaged over the duration of the recording. The main metric used for the frame-based evaluation is the acoustic event error rate (AEER) used in the CLEAR evaluations [19]:

$$AEER = \frac{D + I + S}{N} \tag{1}$$

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 $S = \min\{D, I\}$. Additional metrics include 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:

$$P = \frac{c}{e}, \;\; R = \frac{c}{r}, \;\; F = \frac{2PR}{P + R}. \tag{2}$$

For the event-based metrics, two types of evaluation will take place, an onset-only and an onset-offset-based evaluation. For the onset-only evaluation, each event is considered to be correctly detected if the onset is within a 100ms tolerance. For the onset-offset evaluation, each event is correctly detected if its onset is within a 100ms tolerance and its offset is within 50% range of the ground truth event's offset w.r.t. the duration of the event. Duplicate events are counted as false alarms. The AEER and P-R-F metrics for both the onset-only and the onset-offset cases are utilised.

Finally, in order to ensure that that repetitive events do not dominate the accuracy of an algorithm, class-wise event-based evaluations are also performed. Compared with the event-based evaluation, the AEER and P-R-F metrics will be computed for each class separately within a recording and then averaged across classes. For example, the class-wise F-measure is defined as:

$$F' = \frac{1}{K} \sum_{k} F_k \tag{3}$$

where F_k is the F-measure for events of class k.

4. BASELINE SYSTEMS

4.1. Scene classification

The widespread standard approach to audio classification is the "bag-of-frames" model discussed above. Its modelling assumptions imply among other things that the sequence ordering of frames is ignored [20, 6]. Foote [20] is an early example, comparing MFCC distributions via vector quantisation. Since then, the standard approach to compare distributions is by constructing a Gaussian Mixture Model for each instance or for each class [6, 21].

The MFCC+GMM approach to audio classification is relatively simple, and has been criticised for the assumptions it incurs [22]. However, it is quite widely applicable in a variety of audio classification tasks. Aucouturier and Pachet [6] specifically claim that the MFCC+GMM approach is sufficient for recognising urban soundscapes but not for polyphonic music (due to the importance of temporal structure in music). It has been widely used for scene classification among other recognition tasks, and has served as a basis for further modifications [9]. The model is therefore an ideal baseline for the Scene Classification task.

Code for the bag-of-frames model has previously been made available for Matlab.[4] However, for maximum repro-

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

[4] http://www.jj-aucouturier.info/projects/mir/boflib.zip

ducibility we wished to provide simple and readable code in a widely-used programming language. The Python language is very widely used, freely available on all common platforms, and is notable for its emphasis on producing code that is readable by others. Hence we created a Python script embodying the MFCC+GMM classification workflow, publicly available under an open-source licence,[5] and designed for simplicity and ease of adaptation.

4.2. Event detection

As mentioned in Sec. 2, the NMF framework is a useful one for event detection as it can deal with polyphonic content and the low-rank approximation it provides can efficiently model the underlying spectral characteristics of sources hidden in an acoustic scene. Therefore, we chose to provide an NMF-based baseline system[6] that performs event detection in a supervised manner, using a pre-trained dictionary.

Our algorithm is based on NMF using the β-divergence as a cost function [23]. As a time-frequency representation, we used the constant-Q transform (CQT) with a log-frequency resolution of 60 bins per octave [24]. The training data were normalized to unity variance and NMF with Kullback-Leibler (KL) divergence ($\beta = 1$) was used to learn a set of N bases for each class. The numbers of bases we tested was $5, 8, 10, 12, 15, 20$ and $20i$, the latter corresponding to learning individually one basis per training sample, for all 20 samples. Putting together the sets for all classes we built a "fixed" dictionary of bases used subsequently to factorize the normalized development set audio streams. Afterwards, we summed together the activations per class obtained from the factorization. We tested the use of median filtering for smoothing purposes but this did not improve the classification. Finally a threshold T was chosen to be applied in order to give us the final class activations. The optimal N and T values were chosen empirically by maximizing the F-measure for the two annotations on the development set.

5. RESULTS

The two baseline systems were tested using the public datasets of the challenge described in Sec. 3. In scene classification, where chance performance is 10%, our baseline system attained $52 \pm 13\%$ (95% confidence interval). Table 1 breaks down these results as a confusion matrix. It shows, for example, that *supermarket* was the true class most often mislabelled, most commonly as *openairmarket* or *tubestation*.

Results for the event detection system are shown in Table 2. These include the computed metrics as presented in Sec. 3.3, as well as the optimal system parameters determined separately for both annotations (1-2), calculated as mentioned in Sec.4.2. We found that learning basis vectors from individual sounds resulted in better performance. It is also worth highlighting that the event-based metrics lead to lower reported performance than the frame-based metric.

Finally, not all classes were detected equally well. The F_k was 0% for certain classes, which were: *keyboard*, *keys*, *mouse*, *printer*, and *switch*. All these sounds are characterised by a short-lived and highly transient nature and very low SNR levels that might be potential reasons for failing to be detected by the system. A further set of sounds with an overall poor F_k were: *alert*, *laughter*, and *pageturn*. We believe that the big variation that characterises these sounds could be the reason behind the low performance of the baseline system.

Label	bus	busystreet	office	openairmarket	park	quietstreet	restaurant	supermarket	tube	tubestation
bus	9	-	-	-	-	-	-	-	1	-
busystreet	-	5	-	2	-	-	1	-	-	2
office	-	-	8	-	1	-	-	1	-	-
openairmarket	-	-	-	8	-	-	1	1	-	-
park	-	-	2	1	3	3	-	1	-	-
quietstreet	-	-	-	2	2	4	-	2	-	-
restaurant	-	-	-	2	-	-	3	3	-	2
supermarket	1	-	1	2	1	-	1	2	-	2
tube	-	-	-	-	-	-	2	-	6	2
tubestation	-	-	-	2	-	-	-	2	2	4

Table 1. Confusion matrix for scene classification with baseline MFCC+GMM classifier. Rows are ground-truth labels.

	Evaluation Method		
Metrics / Parameters	Event Based	Class-Wise Event Based	Frame Based
R	16.8	21.7	16.0
P	15.9	11.6	29.1
F-measure	**15.4**	**13.5**	**20.6**
$AEER^*$	2.51	2.94	1.62
Offset R	4.4	7.1	-
Offset P	4.5	3.6	-
Offset F-meas.	4.2	4.0	-
Offset $AEER^*$	2.88	3.37	-
Optimal N	$20i$	$20i$	20-$20i$
Optimal T	750-650	400-700	400-550

* Not measured in (%)

Table 2. Detection accuracy (%) of the NMF system for the Event Detection task for the monophonic Office Live Dataset.

[5] http://code.soundsoftware.ac.uk/projects/smacpy

[6] http://code.soundsoftware.ac.uk/projects/d-case-event

6. CONCLUSIONS

In this paper, we presented a newly launched public evaluation challenge for the classification of acoustic scenes and the detection of acoustic events. We presented the datasets, evaluation metrics, and finally offered evaluation results using an MFCC+GMM system for scene classification and an NMF-based system for event detection. The challenge datasets and the code for both systems are available online and third parties are welcome to use it as the basis for challenge submissions as well as for future research in the CASA field.

Possible extensions for the scene classification system may include a wider set of features, addition of temporal features (such as ΔMFCC) or the use of HMMs to model the various acoustic scenes. For event detection, possible extensions could be to remove or de-emphasize lower frequencies that mainly capture ambient background noise, to try different β values, or to add constrains in the NMF algorithm such as sparsity on the activation matrices. Of course, these baseline systems are just examples, and we welcome approaches to the tasks that differ radically from the baseline systems we have implemented.

At the time of writing, the challenge is still running; results and descriptions of submitted systems will be made available online [5]. In the future, we aim to release detailed challenge results and create a code repository for all open-source submissions, which can serve as a point of reference for the advancement of CASA research.

7. REFERENCES

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

[2] 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.

[3] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A.F. Smeaton, and G. Quénot, "Trecvid 2012–an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proc TRECVID*, 2012.

[4] L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "The Albayzin 2010 language recognition evaluation," in *Proc InterSpeech*, 2011, pp. 28–31.

[5] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. P. Plumbley, "Detection and classification of acoustic scenes and events, an IEEE AASP challenge," Tech. Rep. EECSRR-13-01, Queen Mary University of London, 2013.

[6] 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.

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

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

[9] 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.

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

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

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

[19] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," in *Proc CLEAR*, Southampton, UK, 2007, pp. 311–322.

[20] J. Foote, "Content-based retrieval of music and audio," in *Proc SPIE*, 1997, vol. 3229, pp. 138–147.

[21] S. Wegener, M. Haller, J. J. Burred, T. Sikora, S. Essid, and G. Richard, "On the robustness of audio features for musical instrument classification," in *Proc EUSIPCO*, 2008.

[22] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: how high's the sky?," *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, pp. 1–13, 2004.

[23] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, 2007.

[24] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc SMC*, Barcelona, Spain, July 2010, pp. 3–64.