DETECTION AND CLASSIFICATION OF ACOUSTIC SCENES AND EVENTS: AN *IEEE AASP* CHALLENGE

Dimitrios Giannoulis[1], *Emmanouil Benetos*[2], *Dan Stowell*[1], *Mathias Rossignol*[3], *Mathieu Lagrange*[3] and *Mark D. Plumbley*[1]

[1] Centre for Digital Music, School of EECS, Queen Mary University of London, London, UK
[2] Department of Computer Science, City University London, London, UK.
[3] Analysis/Synthesis Team, IRCAM, CNRS-STMS, Paris, France.

ABSTRACT

This paper describes a newly-launched public evaluation challenge on acoustic scene classification and detection of sound events within a scene. Systems dealing with such tasks are far from exhibiting human-like performance and robustness. Undermining factors are numerous: the extreme variability of sources of interest possibly interfering, the presence of complex background noise as well as room effects like reverberation. The proposed challenge is an attempt to help the research community move forward in defining and studying the aforementioned tasks. Apart from the challenge description, this paper provides an overview of systems submitted to the challenge as well as a detailed evaluation of the results achieved by those systems.

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

1. INTRODUCTION

Over the last few years, there has been an increased interest in the speech and audio processing community in code dissemination and reproducibility of results as a means to improve the quality and relevance of published results. This can be attributed to accumulating evidence of the benefits of performing research with reproducibility in mind and making well-documented code and data publicly available [1, 2]. Public evaluation of proposed methods, especially if accompanied with open-source submissions is a key component in the move towards this reproducibility. It can serve as a reference point for the performance of proposed methods and can also be used for studying performance improvements throughout the years. Numerous initiatives have reached maturity, for example the SiSEC evaluation for signal separation [3], the MIREX competition for music information retrieval [4] and the CHiME speech separation and recognition challenge [5]. The research problems related with these evaluations are well-defined and have their own performance metrics established. However, for researchers working on modeling and classifiction of acoustic scenes, containing non-speech and non-music, and detecting sound events within a scene, there is not yet a coordinated established international challenge in this area, with the exception of the now discontinued CLEAR evaluations [6] funded by the CHIL project and the Multimedia Event Detection of the TRECVID video retrieval evaluations, where the focus is on audiovisual, multi-modal event detection in video recordings [7].

In this paper, we describe a newly-launched public evaluation challenge on acoustic scene classification and event detection, both for monophonic and polyphonic audio [8]. In Section 2, we present the datasets that were created for the challenge, as well as the employed evaluation metrics. Participating systems are then outlined in Section 3, and evaluation results are presented and discussed in Section 4.

2. CHALLENGE DESCRIPTION

Acoustic scene classification and *detection of sound events within a scene* are well defined engineering tasks that both fall under the "umbrella" of computational auditory scene analysis (CASA) [9]. The first task aims to characterize the acoustic environment of an audio stream by providing a semantic label to it [10]. The second one aims to label temporal regions within the audio, within which a specific event class is active, by estimating the start and end time of each event and if necessary (i.e. for audio extraction purposes) separate it from other overlapping events.

The present challenge consists of a set of three subtasks. The first one addresses the problem of identifying and classifying acoustic scenes or soundscapes. The other two subtasks address the problem of identifying individual sound events that are prominent in an acoustic scene: one focuses on monophonic event detection without overlapping sounds and the other focuses on polyphonic scenarios. The polyphonic case could be considered more interesting, as in realistic everyday scenarios most of the sounds that reach our ears tend to stem from a multitude of sources, but at the same time it consists of a much more challenging problem. More details about the proposed tasks along with baseline results can be found in [8, 11].

2.1. Datasets

Each of the tasks is accompanied by its own dataset. The datasets for Scene Classification (SC) consists of two equally proportioned parts each made up of ten 30 seconds recordings for each scene (class), for a total of 100 recordings per dataset. One part has been made publicly available [8] and serves as the development set for participants to investigate the performance of their system, whereas the other is kept private and used for a train/test (K-fold) evaluation. The two datasets span 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*.

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.

Participants	**Code**	**Method**	**Lang**
Chum et al.	CHR	Various features at 2 frame sizes, classified either: (a) per-frame SVM + majority voting; (b) HMM	Matlab
Elizalde	ELF	Concatenation of 4 different mono mixdowns; “i-vector” analysis of MFCCs, classified by pLDA	Matlab
Geiger et al.	GSR	Diverse features, classified within 4-second windows using SVM, then majority voting	Weka/ HTK
Krijnders and ten Holt	KH	“Cochleogram” representation, analysed for tonelikeness in each t-f bin, classified by SVM	Python
Li et al.	LTT	Wavelets, MFCCs and others, classified in 5-second windows by treebagger, majority voting	Matlab
Nam et al.	NHL	Feature learning by sparse RBM, then event detection and max-pooling, classified by SVM	Matlab
Nogueira et al.	NR1	MFCCs + MFCC temporal modulations + event density estimation + binaural modelling features, feature selection, classified by SVM	Matlab
Olivetti	OE	Normalised compression distance (Vorbis), Euclidean embedding, classified by Random Forest	Python
Patil and Elhilali	PE	Auditory representation analysed for spectrotemporal modulations, classified within one-second windows using SVM, then weighted combination of decision probabilities	Matlab
Rakotomamonjy and Gasso	RG	Computer vision features (histogram of oriented gradient) applied to constant-Q spectrogram, classified by SVM	Matlab
Roma et al.	RNH	Recurrence Quantification Analysis applied to MFCC time-series, classified by SVM	Matlab
Baseline		MFCCs, classified with a bag-of-frames approach	Python

Table 1: Summary of submitted scene classification systems.

Participants	**Code**	**Method**	**Lang**
Chauhan et al.	CPS	Feature extraction - Segmentation - Likelihood ratio test classification	Matlab
Diment et al.	DHV	MFCCs (features) - HMMs (detection)	Matlab
Gemmeke et al.	GVV	NMF (detection) - HMMs (postprocessing)	Matlab
Niessen et al.	NVM	Hierarchical HMMs + Random Forests (classification) - Meta-classification	Matlab
Nogueira et al.	NR2	MFCCs (features) - SVMs (classification)	Matlab
Schröder et al.	SCS	Gabor filterbank features - HMMs (classification)	Matlab
Vuegen et al.	VVK	MFCCs (features) - GMMs (detection)	Matlab
Baseline		NMF with pre-extracted bases (detection)	Matlab

Table 2: Summary of submitted event detection systems.

These recordings were made with a set of Soundman OKM II binaural microphones. These microphones imitate a pair of in-ear headphones that the user can wear for added portability and subtlety. Furthermore, the data carries also binaural information about the sound that could be utilized as cues for the sound event and scene detection or simply be ignored by adding the two channels together in order to obtain a mono recording.

The datasets for event detection were built from audio collected in office environments because of the interest of such audio to certain applications such as audio-conferencing systems etc. Two event detection tasks are proposed, a monophonic task denoted as Office Live (OL) and a polyphonic task denoted as Office Synthetic (OS). Polyphonic data for the OS task was created using a scene synthesizer, concatenating recordings of isolated events. Each dataset consists of three subsets (a training, a development and a testing dataset). The training set contains instantiations of individual events for every class and is shared between the OL and OS tasks to allow for single training of event detection systems. The development and testing datasets consist of roughly 1 minute long scripted recordings of everyday audio events in a number of office environments (different size and absorbing quality rooms, different number of people in the room and varying noise level). 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 (4-channel), with a high-quality Soundfield SPS422B microphone system, 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 to extend the challenge to full B-format in future if there is demand.

More details about the creation of the datasets, the annotation process and the audio recording process can be found in [8].

2.2. Evaluation Metrics

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

For the event detection tasks, in order to provide a thorough assessment of the various systems, three types of evaluations take place, namely a frame-based, event-based, and class-wise event-based evaluation. The main metrics used for each evaluation type are the F-measure (F) and the acoustic event error rate ($AEER$) as described in [8]. For the event-based and class-wise event-based metrics, two types of evaluation will take place, an onset-only and an onset-offset-based evaluation. Results to onset-based metrics (denoted without any subscript) and onset-offset-based metrics (denoted as F_{offset} and $AEER_{offset}$). For a complete and analytic description of the evaluation metrics employed the reader is referred to [8, 11].

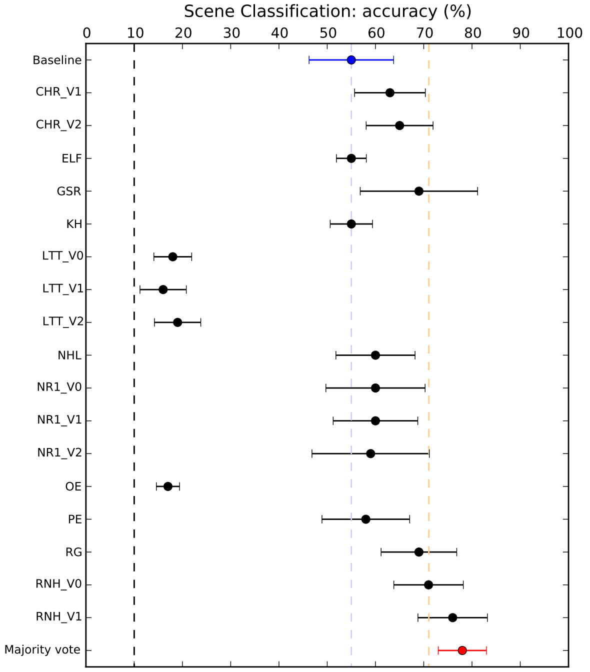

Figure 1: Classification accuracy(%) for the SC task. Plot shows mean over 5-fold cross-validation with 95% confidence intervals. Dashed lines indicate (left to right): chance performance (black); baseline system performance (light blue); mean accuracy of human listener (orange). "Majority vote" is a meta-classifier using the majority decision over all submissions.

3. SUBMITTED SYSTEMS

Overall, 11 systems were submitted to the SC task, 7 systems were submitted to the OL task, and 3 systems to the OS task. Variants for each system were allowed that increased the total number and variety considerably. The systems submitted for the scene classification and event detection tasks are listed in Tables 1 and 2, respectively, along with a short description of each system and the programming language in which it was written.

Apart from the submitted systems, performance on the test sets is also reported for baseline systems for the two tasks. These systems were made publicly available as open source software [11].

4. CHALLENGE RESULTS

Results were computed by running all the submitted systems on the held-back testing datasets and computing the metrics as in Sec. 2.2. Figure 1 shows the overall performance of submitted systems for the scene classification task. Most systems were able to outperform the baseline system, and some matched or even slightly excelled the mean accuracy we found in human listeners (71%; results in preparation). The strongest performers are notably diverse in their choice of features and their use of temporal information, though often using SVMs for classification. Two submissions achieved good results on the development data but not on our held-out test data.

Table 3 shows a confusion matrix for the scene labels as round percentages of the sum of all confusion matrices for all submissions. Confusions are mostly concentrated over classes that share some acoustical properties such as park/quietstreet and tube/tubestation.

Label	bus	busystreet	office	openairmarket	park	quietstreet	restaurant	supermarket	tube	tubestation
bus	**81**	3	0	4	1	0	0	4	6	2
busystreet	1	**69**	14	2	1	2	1	3	3	5
office	1	0	**55**	13	9	12	4	3	1	3
openairmarket	1	2	0	**59**	13	0	9	12	3	2
park	1	1	8	3	**51**	29	3	2	1	1
quietstreet	0	5	4	3	29	**43**	9	5	0	1
restaurant	1	1	0	16	5	0	**53**	21	2	3
supermarket	6	5	6	6	4	7	10	**42**	7	7
tube	7	7	1	1	2	2	5	3	**44**	28
tubestation	5	16	1	4	1	2	3	8	19	**41**

Table 3: Aggregate confusion matrix for scene classification across all submissions. Rows are ground truth, columns the inferred labels. Values are expressed as percentages rounded to the nearest integer.

For the event detection OL and OS tasks, results are summarized in Tables 4 and 5, respectively. The baseline was outperformed by most systems for these tasks too. The best performance for the OL task using all types of metrics is achieved by the SCS submission, which used a Gabor filterbank feature extraction step with by 2-layer hidden Markov models (HMMs) for classifying events, followed by the NVM submission, which used a meta-classifier combining hierarchical HMMs and random forests. For the OS task, the best performance in terms of F-measure is achieved by the DHV system, which used an iterative scheme with HMMs. It should also be noted that submitted systems performed better with lower polyphony, with the exception of the DHV system, which had better performance with higher polyphony levels. As expected, the onset-offset evaluation produced worse results compared to onset-only evaluation for both tasks, although the performance difference is rather small. This may be explained by the percussive nature of most events.

The challenge website [8] gives detailed system descriptions and extensive results, analytic breakdown of performance per system, as well as further error analysis.

5. CONCLUSIONS

In this paper we presented a challenge on the detection and classification of acoustic scenes and events. We ran a scene classification (SC) challenge, and two event detection and classification challenges: office live (OL) and office synthetic (OS). Our goal was 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 results enable us to draw some interesting conclusions about the different problems. For scene classification, we found that although simple systems can do relatively well, the improvement that more complex systems achieve can bring performance to the levels achieved by human listeners. For event detection, which is a more challenging task, performance is much worse although we have not performed a direct comparison with human listeners at present. For the monophonic case, systems are able to achieve satisfactory performance with scope for improvement. For the polyphonic case, the task of recognising individual potentially overlapping sounds becomes significantly challenging and the performance of systems that are even prepared to deal with polyphonic content

	Evaluation Method									
	Event-Based				Class-Wise Event-Based				Frame-Based	
System	F (%)	F_{offset} (%)	$AEER$	$AEER_{offset}$	F (%)	F_{offset} (%)	$AEER$	$AEER_{offset}$	F (%)	$AEER$
CPS	2.23	1.65	2.285	2.301	0.65	0.49	1.872	1.891	3.82	2.116
DHV	26.67	22.43	2.519	2.676	30.72	25.29	2.182	2.370	26.0	3.128
GVV	15.52	13.46	1.779	1.831	13.21	12.03	1.556	**1.606**	31.94	1.084
NVM_1	32.57	24.95	1.864	2.095	29.37	21.80	1.639	1.899	40.85	1.115
NVM_2	34.16	26.28	1.852	2.095	33.05	24.88	1.602	1.877	42.76	1.102
NVM_3	34.51	27.01	1.827	2.052	33.52	24.65	1.575	1.846	45.50	1.212
NVM_4	30.47	24.68	1.906	2.083	28.17	21.62	1.650	1.849	42.86	1.360
NR2	19.21	15.26	3.076	3.244	21.54	17.64	2.857	3.010	34.66	1.885
SCS_1	39.47	36.74	1.669	1.749	36.33	34.20	1.579	1.677	53.02	1.167
SCS_2	**45.17**	**41.06**	**1.601**	**1.727**	**41.51**	**38.32**	**1.511**	1.646	**61.52**	1.016
VVK	30.77	25.40	2.054	2.224	24.55	20.36	1.762	1.949	43.42	**1.001**
Baseline	7.38	1.58	5.900	6.318	9.00	1.86	5.960	6.462	10.72	2.590

Table 4: Evaluation metrics for the participating systems for the (monophonic) Office Live Event Detection task.

	Event-Based				Class-Wise Event-Based				Frame-Based	
System	F (%)	F_{offset} (%)	$AEER$	$AEER_{offset}$	F (%)	F_{offset} (%)	$AEER$	$AEER_{offset}$	F (%)	$AEER$
DHV	**8.45**	6.18	4.741	4.860	**9.73**	**7.58**	4.028	4.147	**13.08**	8.426
GVV	7.69	**7.33**	1.913	1.920	6.69	6.51	1.584	1.591	10.30	**1.553**
VVK	5.80	5.28	**1.885**	**1.895**	5.10	4.77	**1.436**	**1.445**	5.77	2.106
Baseline	4.98	0.24	6.507	6.895	6.69	0.18	5.389	5.782	6.88	3.047

Table 5: Evaluation metrics for the participating systems for the (polyphonic) Office Synthetic Event Detection task.

falls dramatically. More details for all the submitted systems can be found on the challenge website in [8].

At this point, we have just completed running the challenge. For future work, we will consider producing a detailed performance evaluation, creating a code repository, releasing test sets, doing a B-format challenge, running the challenge again or doing a challenge on world synthetic sounds (WS) as proposed in [8].

6. ACKNOWLEDGMENTS

The authors would like to thank all of the contributors, for taking part in the challenge and participating in email discussions as we were developing the challenge, and Daniele Barchiesi, for his contribution in running the evaluation of the submitted systems and providing useful feedback.

7. REFERENCES

[1] P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible research in signal processing," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 37–47, 2009.

[2] J. Kovacevic, "How to encourage and publish reproducible research," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2007, pp. 1273–1276.

[3] S. Araki, F. Nesta, E. Vincent, Z. Koldovskỳ, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011)," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 414–422.

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

[5] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, May 2012.

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

[8] 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," http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/, Queen Mary University of London, Tech. Rep. EECSRR-13-01, 2013.

[9] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEE Press, 2006.

[10] 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," *J. Acoust. Soc. of America*, vol. 122, no. 2, pp. 881–891, 2007.

[11] 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 *European Signal Processing Conf.*, 2013.