IEEE AASP Challenge

Detection and Classification of Acoustic Scenes and Events

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

I. Problem Formulation

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 [3] 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 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 current proposal.

Finally, one public evaluation that is related to the proposed challenge took place in 2006 and 2007, as part of the CLEAR evaluations [8], 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 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

[†] 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

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. Challenge Description

Two closely related tasks in computational auditory scene analysis (CASA) are *acoustic scene classification* and *detection of sound events within a scene* [1]. A system involved in the first task has as a goal to characterise or "label" the environment in which the audio was recorded [6], 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.

The aim of the proposed challenge would be 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 will address the problem of identifying and classifying acoustic scenes and soundscapes. The second challenge will address the problem of identifying individual sound events that are prominent in an acoustic scene. Two distinct experiments will 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.

III. Evaluation Data

A. Dataset

There will be four datasets overall, one for scene classification and three for event detection. The first one for the scene classification (SC) challenge will consist of 30 sec recordings of various acoustic scenes. The dataset will consist 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 will be sent out to the participants to build up and investigate the performance of their system and the other will be kept secret and used for the train/test Scene Classification task. Scenes would be:

- busy street
- quiet street
- supermarket/store
- restaurant
- office
- park
- bus
- tube/metro
- tubestation
- open market

The second dataset will consist of three subsets (a training, a development and a testing dataset). The training set[1] will contain instantiations of individual events for every class. The development (validation) and testing dataset, denoted as office live (OL), will consist 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 for these recordings will be annotated and they will 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

[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

There will be two different annotations that will be released, each made from a different person and both examined for consistency and 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 will help 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. The testing is planned to be carried out on both. The train set will include 24 recordings of individual sounds per class, followed by annotations of their onset and offset in sec. The development set will include 3 recordings of a series of events from one of the office environments. These recordings will also be accompanied by annotation of the events' onsets and offsets. The third set that will not be released will have recordings of sound events made in all office-environments, excluding the one used for the development set.

The third dataset will contain artificially sequenced sounds provided by the Analysis-Synthesis team of IRCAM, termed Office Synthetic (OS). The data for the OS task will consist of three subsets like the previous task. The training dataset will consist of audio recordings of individual events and will be identical to the one for the realistic task.The development and testing datasets will 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 will be recorded by QMUL specifically for this challenge, confidentiality is ensured. 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 will be most accurate even for polyphonic mixtures with lots of overlaps among different sounds. We would 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 4th and last dataset will contain artificially sequenced sounds provided by the Analysis-Synthesis team of IRCAM, termed World synthetic (WS). The training set will consist of audio recordings of a wide range of individual events from a predefined set of events taken from the FreeSound database (http://www.freesound.org/). As in the OS task, the development and testing datasets will be built from different recordings of the same type of events (also taken from the FreeSound database). Confidentiality is not ensured in this task, as we need to trust the participants not to train and develop their system using the FS samples used for building the testing set. For that reason the Freesound IDs of the samples used for building the testing set will be provided to the participants. The aim of this subtask is to study the behavior of tested algorithms when facing a more diverse type of events and backgrounds. The candidate event classes will include for sea scenes: waves, birds, splashing, motor boat, fog horn, footsteps (sand, gravel), kids playing, wind gusts, and sail noises. For forest scenes: wind gusts, rustling leaves, footsteps (human, animals), bird calls, insects, mammal calls, hunting sounds, and branches breaking.

Note: All datasets for the challenge will be released under a creative commons (CC BY) license[2].

B. Recording Equipment

The Centre for Digital Music at Queen Mary University of London has 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 [10], 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 [11], 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) citeCheng. Therefore, the resulting data carry also binaural information about the sound that could additionally be utilized as cues for sound event and scene detection from audio or simply be ignored entirely by adding the two channels together in order to obtain a mono recording.

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

The sound files for the 1st task (acoustic 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 will also be released together with the stereo versions but the challenge will be run on stereo and not the B-format. The participants for the challenge will 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 (exluding task 4 where recording conditions were not controlled).

IV. Metrics

A. Scene Classification

For classifying acoustic scenes, the output of each run for a single file would only contain the class label. As in the MIREX train/test tasks [2], the metrics that will be computed will be 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 will be evaluated using 5-fold cross validation.

B. Event Detection

For event detection, three types of evaluations will 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 will be 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 will be performed using a 10ms step. The main metric utilised for the frame-based evaluation would be 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 can be 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 will be considered as false alarms.

Finally, a class-wise event-based evaluation will also take place, in order to ensure that that repetitive events do not dominate the accuracy of an algorithm. The output of the algorithm will be the same as in the event-based evaluation, but in this fase the AEER and P-R-F metrics will be 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.

V. Website

At present, we are hosting a webpage for the task that can be reached at: http://www.elec.qmul.ac.uk/digitalmusic/sceneseventschallenge/. The webpage includes a brief description of the challenge and some sample recordings. In the near future and given that the task makes it to the next step we will host further details about the metrics, the datasets and the overall evaluation task so as to fuel a discussion between groups and researchers that have expressed their interest in participating in the challenge and help better define some aspects of the task.

The challenge also includes a dedicated mailing list for facilitating discussion: http://www.eecs.qmul.ac.uk/mailman/listinfo/aasp-challenge.

VI. Contacted Groups

A number of researchers with active research on the field in the past years has been contacted and informed about the proposed challenge with most of them having already expressed an interest in participating. If anything, this was to gather an initial expression of interest from various groups and researchers to take part in the challenges discussion as well as in all, or parts, of the evaluation. During the discussion for the challenge and the call for participation we encourage researchers to express their interest on the challenge and work together in formalising all the aspects of the challenge.

VII. 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).

[1] We have already contacted the WASPAA 2013 chairs and they have agreed to assign a time slot for presenting the challenge and its results. Since the workshop schedule is not defined yet we would have to finalise the exact form of the session in a later date.

References

[1] D.L. Wang and G. J. Brown (Eds), “Computational auditory scene analysis: Principles, algorithms and applications,” IEEE Press/Wiley-Interscience, 2006.

[2] MIREX Campaign, http://www.music-ir.org/mirex

[3] SiSEC Evaluation, http://sisec.wiki.irisa.fr

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

[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] 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, No. 2, pp. 881-891, 2007.

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

[8] CLEAR Evaluation, http://clear-evaluation.org/

[9] Corey I. Cheng, and Gregory H. Wakefield, “Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space,” Journal Audio Eng Soc, Vol 49, No 4, 2001 April.

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

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