Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge

Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, Mark D. Plumbley

Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge

Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, *Member, IEEE,* Peter Foster, *Member, IEEE,* Mathieu Lagrange, Tuomas Virtanen, *Senior Member, IEEE,* and Mark D. Plumbley, *Fellow, IEEE*

***Abstract*—Public evaluation campaigns and datasets promote active development in target research areas, allowing direct comparison of algorithms. The second edition of the challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016) has offered such an opportunity for development of state-of-the-art methods, and succeeded in drawing together a large number of participants from academic and industrial backgrounds. In this paper, we report on the tasks and outcomes of the DCASE 2016 challenge. The challenge comprised four tasks: acoustic scene classification, sound event detection in synthetic audio, sound event detection in real-life audio, and domestic audio tagging. We present in detail each task and analyse the submitted systems in terms of design and performance. We observe the emergence of deep learning as the most popular classification method, replacing the traditional approaches based on Gaussian mixture models and support vector machines. By contrast, feature representations have not changed substantially throughout the years, as mel frequency-based representations predominate in all tasks. The datasets created for and used in DCASE 2016 are publicly available and are a valuable resource for further research.**

***Index Terms*—Acoustic scene classification, audio datasets, pattern recognition, sound event detection**

I. Introduction

Environmental sound classification and detection is a rapidly developing research area. Its growth has been stimulated by emerging public evaluation campaigns and datasets promoting active development in areas like automatic classification of acoustic scenes and automatic detection and classification of sound events. The series of challenges on Detection and Classification of Acoustic Scenes and Events (DCASE) provides a great opportunity for development and comparison of state-of-the-art methods, by offering a set of tasks with corresponding datasets, metrics and evaluation frameworks for specific topics within this research field.

Manuscript received Month, Day, 2017; revised Month, Day, 2017. AM, TH and TV received funding from the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND. EB is supported by a UK RAEng Research Fellowship (RF/128). MDP is partly supported by a Grant (EP/N014111/1) from the UK Engineering and Physical Sciences Research Council (EPSRC).

A. Mesaros, T. Heittola and T. Virtanen are with the Department of Signal Processing, Tampere University of Technology, 33720, Finland, e-mail: {annamaria.mesaros, toni.heittola, tuomas.virtanen}@tut.fi

E. Benetos and P. Foster are with the Centre for Digital Music, Queen Mary University of London, London E1 4NS, U.K., email: {emmanouil.benetos, p.a.foster}@qmul.ac.uk

M. Lagrange is with the ADTSI, IRCCYN, Ecole Centrale de Nantes, Nantes 44321, France.

M. D. Plumbley is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K., email: m.plumbley@surrey.ac.uk

Evaluation campaigns are common in many research areas and play an important role in advancing research and algorithm development. In the broad field of audio processing, automatic speech recognition evaluations have a long history [1], while the Music Information Retrieval Evaluation eXchange (MIREX) [2] has been running yearly for over a decade already. From neighboring research areas, the TRECVid Multimedia Event Detection (MED) evaluation track [3] that deals with detecting user defined events in videos, includes and encourages use of audio information for detection. Related public evaluation campaigns also include SiSEC challenge on signal separation [4] and the REVERB challenge on reverberant speech processing research [5]. Over the years, the proposed evaluation tasks in these campaigns have grown in data size, data complexity and task difficulty. In addition, evaluation campaigns that deal with more specialized topics have also appeared, for example detection of birds in audio [6].

Research in environmental sound classification and detection is part of *computational auditory scene analysis*, and is currently receiving large amounts of interest within the audio research community, manifested through special issues and sessions in related journal and conferences. The high volume of recent publications on such topics is fueled by interest in context awareness, content-based information processing of continuously growing amounts of audio material, and not least by the development of strong computational methods based on deep learning architectures. Two main research directions are evident within the computational auditory scene analysis field: acoustic scene classification as a general environment recognition problem, and sound events classification or detection as a more detailed attempt at describing the environment through the sounds encountered in it.

Acoustic scene classification is based on the premise that it is possible to provide a textual label as a general characterization of a location or situation, which is assumed to be distinguishable from others based on its general acoustic properties. The problem is typically framed as supervised classification, and often involves a relatively small number of classes. A thorough review of features and classifiers used for acoustic scene classification is presented in [7], presenting in detail the approaches submitted for DCASE 2013. Existing approaches often include use of mel-frequency cepstral coefficients and other low level spectral descriptors [8], [9] or more specialized features such as histograms of sound events [10] or histogram of gradients learned from time-frequency representations [11]. On the acoustic modeling

aspect, methods range from classical statistical models like hidden Markov models (HMMs) [8], Gaussian mixture models (GMMs) [9] or support vector machines (SVMs) [11], to more recently developed methods using deep learning that have high computational complexity in training and often have a large number of parameters [12].

Sound event detection and classification are based on the premise that sounds produced from the same source or through the same physical process can be grouped into a category, and can be distinguished from sounds originating from different sources or through different processes. In existing literature there is often not a clear distinction between detection and classification, with many early works dealing only with classification of isolated sounds. Hereafter we refer to sound event detection within an audio segment as classifying the sound into a category and locating it within the audio in terms of onset and offset relative to the entire duration. Simplified scenarios include having a single sound event per audio segment [13] or a sequence of non-overlapping sound events as the Office Live task in DCASE 2013 [6]. The most complex variant of sound event detection, referred to as *polyphonic*, involves detection of overlapping sound events. Often based on mel-scale spectral representations of the signal for features, employed methods for sound event detection include HMMs [14], NMF [15]–[17], and recently a variety of temporally constrained deep learning methods such as convolutional neural networks (CNNs) [18]–[20] and long short-term memory (LSTM) [21], [22].

As an alternative to acoustic scene classification and event detection, we may attempt to characterise an audio segment by assigning to it one or more labels, where each label indicates the presence of a particular acoustic event class in the audio segment without the need to locate the event. Thus formulating *audio tagging* as a multi-label classification task, we may consider the particular case where each training instance is an audio segment with a set of assigned labels. Since the labels provide no indication about onset and duration of acoustic events, we may consider such data weakly labeled. Whereas audio tagging has been widely applied for analyzing musical recordings [23]–[29], environmental audio tagging remains comparatively unexplored. In current studies, methods investigated include GMMs [30]–[32], SVMs combined with multiple instance learning [33], unsupervised feature learning [34], [35] and CNNs [36].

Interest for automatic environmental sound recognition has seen significant growth recently; however, in contrast to resources supporting speech or music research, databases containing environmental sounds are not easily accessible. Recently AudioSet, a large scale dataset for environmental sound research, has been made available by Google [37], containing tags for 10-second audio segments within YouTube videos; its usability in research tasks is yet to be established. Currently available literature on environmental sound recognition uses in-house datasets, making it difficult to have a fair comparison of the methods. An important step towards improving this situation was the first Detection and Classification of Acoustic Scenes and Events (DCASE) challenge organized in 2013 with purpose-built datasets. Even though the amount of data offered was rather small, the challenge introduced public evaluations of everyday sounds. DCASE 2013 was a successful first edition, covering two tasks and attracting submissions from 18 international teams, that concluded with a special session at WASPAA 2013. Thereafter, many other special sessions on environmental sound classification were organized at different conferences, marking a clear boost in research community interest in the topic.

DCASE 2016 was the second edition of the challenge, bringing the tasks closer to real life applications by using complex audio recorded in everyday life, and providing larger amounts of data for the tasks. It was organized as an IEEE Audio and Acoustic Signal Processing Technical Committee challenge, like DCASE 2013, and had a very high amount of participants overall, with four times more submissions than the first challenge. Challenge results were presented during a dedicated one day workshop. Participants came from both academia and industry, showing ongoing research and active development on both sides.

In this paper we present the tasks and outcome of the DCASE 2016 challenge, reporting advances made in the last three years. In Section II we present the DCASE 2016 Challenge organization details, timeline and tasks. We proceed with the detailed presentation of each task in Sections III–VI. For each task, we provide the definition, dataset description and experimental setup, the metrics used for evaluation of the methods, the baseline system provided to the participants as reference performance, and the analysis of submitted systems and results. Finally, Section VIII presents conclusions and provides suggestions on future work and keeping DCASE active.

II. Challenge tasks and timeline

Building on the experience from the first challenge, the tasks for DCASE 2016 were designed to improve upon those in DCASE 2013. The tasks were: Task 1 - Acoustic scene classification, Task 2 - Sound event detection in synthetic audio, Task 3 - Sound event detection in real-life audio, and Task 4 - Domestic audio tagging. Notably, Task 1 was defined the same way as in DCASE 2013, but with a new and much larger dataset, and Task 2 also considered the overlap between sound events to be detected. In addition, Task 3 was introduced to bring the challenge closer to real world applications, and Task 4 was introduced to provide a multi-label classification task.

A key difference between DCASE 2013 and 2016 is that the former asked from participating teams to submit source code for the developed systems, which was run and evaluated by the challenge organizers, thus evaluation data were not released to participants at the time. The 2016 version of the challenge instead released the evaluation data (without reference annotations) to participants, who submitted their system outputs to the challenge organizers for computation of performance metrics.

The advantage of releasing evaluation data instead of requiring source code is that it avoids potential software or output formatting incompatibilities arising from having to execute code collected from participants. For DCASE 2013,

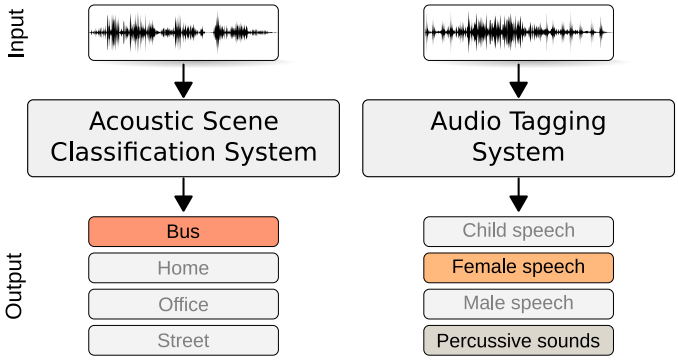

Fig. 1. Acoustic scene classification and audio tagging.

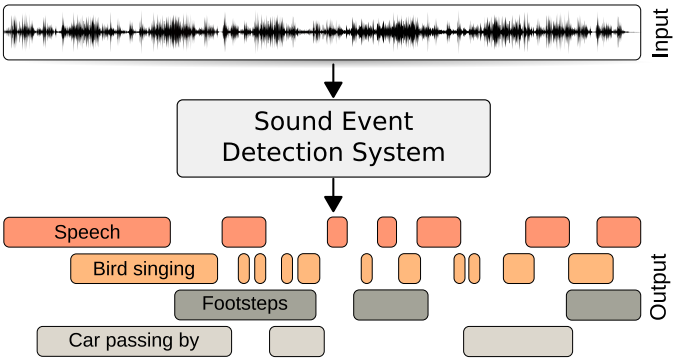

Fig. 2. Sound event detection: finding temporal positions and textual labels for sound events in an audio example.

the organizers indeed reported various software issues with libraries, Linux/Windows differences, formatting and bugs in the submitted code, which are all avoided by requiring submission of system output [38]. Given the substantially increased number of submissions for DCASE 2016, running code for all submissions would require a substantial amount of both computational and human resources. Also, requiring submission of source code may deter participants that are not comfortable or confident in their software development skills. For example the MIREX public evaluation campaign requires source code with strict rules for running it [39]; MIREX traditionally does not have many participants per task, but even so there is an imposed execution time limit, and there are cases in which the allocated execution time is exceeded. On the other hand, submission of system output to challenge organisers does not allow for a execution time analysis of submitted systems, and this practice neither actively promote good software engineering practices nor software sustainability and reproducibility. There are also potential issues with releasing datasets to participants: e.g. for MIREX several datasets are copyright-restricted (which is why they are not shared with participants), as well as on re-using datasets for several editions of a challenge.

A. Task descriptions

Acoustic Scene Classification is an audio classification problem that carries broad interest due to the development of context-aware devices and applications. It is a straightforward multi-class supervised classification problem in which the categories for classification are labels describing the acoustic scene. Figure 1 illustrates in the left panel the way the task is defined: for each audio example, the system must provide a single label; the system is trained using audio data labeled in the same way, with a single label per audio example.

Sound event detection is defined as the task of finding individual sound events in a test audio example by indicating onset, offset and textual labels for each sound event instance, as illustrated in Fig. 2. The sound event classes are predefined, making it a supervised learning task, with training data available for all classes. There were two sound event detection tasks in DCASE 2016, one using synthetic data generated from isolated sound event examples, for which training data were available as isolated sound examples, and the other using recordings of everyday scenes, for which training data containing overlapping sounds was provided, with manually annotated reference similar to the system output illustration. Use of synthetic data allows control over the number and relative levels of overlapping sounds, mixtures containing balanced classes and computation of performance metrics with reliable reference annotations. Real-life audio is more challenging, since real-life sound event classes are often unbalanced: some sound events may be arbitrarily rare, and manual annotations are subjective in both label and onset/offset positioning.

Audio tagging is defined as a multi-label classification problem, in which each possible label corresponds to a class of sound events which may occur in the acoustic scene, as illustrated in Fig. 1. When applied to short audio chunks, audio tagging can be viewed as a coarse-grained variant of sound event detection, where for each audio chunk the presence of a given label informs about whether events of a particular class occur in the chunk. Whereas the onset and duration prediction that we obtain in sound event detection is not requested in audio tagging, the temporal resolution imposed by the audio chunk size may nonetheless be sufficient for typical applications such as human activity monitoring, where predicting precise event boundaries is secondary to characterizing the acoustic scene. A potential practical benefit of audio tagging is the straightforward manual annotation process, which does not necessitate recording event boundaries. The task thus raises an interesting technical challenge, namely how to learn from such weakly labeled data.

As presented in detail in the following sections, each task carries its own distinct objective. Thus, the design of the datasets and the way the metrics are computed may differ, leading to the use of specific statistical significance evaluation procedures for each task.

B. Challenge timeline and participants

Organization of the challenge started in summer 2015 by planning the tasks, data recording and annotation process, converging to the definition of the four tasks. Once the tasks and evaluation procedure were agreed on, the challenge was announced to the community, and the organization procedure started. Table I lists the challenge timeline.

TABLE I
DCASE 2016 CHALLENGE SCHEDULE.

Phase	Time
Challenge announcement	June 2015
Data recording and annotation for tasks 1 and 3	June-Dec 2015
Definition of tasks and evaluation procedure	Sept-Nov 2015
Publication of challenge tasks	Dec 2015
Publication of development datasets and baseline systems	Jan 2016
Publication of evaluation datatsets	Apr 2016
Submission deadline	June 2016
Publication of results	Aug 2016

TABLE II
DCASE 2016 CHALLENGE SUBMISSION STATISTICS.

Task	Submissions	Teams	Authors
Acoustic Scene Classification	48	34	113
Sound Ev. Det. in Synth. Audio	10	9	37
Sound Ev. Det. in Real-Life Audio	16	12	45
Domestic Audio Tagging	8	7	23

The 2016 challenge attracted a substantially higher number of participants than the previous challenge, with a total of 82 submissions for the 4 tasks, with 48 tasks submitted for Task 1 (Acoustic Scene Classification). In comparison, DCASE 2013 comprised a total of 24 submissions from 18 teams. Table II lists statistics on the number of submissions and participants, while more detailed information for each task will be presented in the following sections.

III. ACOUSTIC SCENE CLASSIFICATION

The goal of acoustic scene classification is to classify a test recording into one of predefined classes that characterizes the environment in which it was recorded, such as ”park”, ”bus” ”home”, ”office”.

A. Dataset and experimental setup

The task used the TUT Acoustic Scenes 2016 dataset [40], consisting of recordings from 15 acoustic scenes: lakeside beach, bus, café/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, urban park, residential area, train, and tramway. The acoustic scene categories were selected while planning the data recording procedure. All data were recorded in Finland. To obtain high acoustic variability for all acoustic scene categories, each recording was made in a different location: different streets, different parks, different homes. There are 15-18 locations for each acoustic scene category except office, for which there are only 13. For each recording location, a 3-5 minute long audio recording was captured. Recordings were made using a Soundman OKM II Klassik/studio A3, electret binaural microphone worn in the ears, and a Roland Edirol R-09 recorder using 44.1 kHz sampling rate and 24 bit resolution. All recorded audio material was then cut into segments of 30 seconds length.

The dataset was split into a development set and evaluation set, with the evaluation set consisting of approximately 30% of the total amount. The development set was further partitioned into four folds of training and testing sets to be used for cross-validation during system development. For each acoustic scene, 78 segments were included in the development set and 26 segments were kept for evaluation. The partitioning of the data was based on the location of the original recordings such that all segments obtained from the same original recording were included into a single subset – either development or evaluation, and within the development set into either the training or testing subset.

B. Baseline system and evaluation metric

The baseline system provided for the task [40] consists of a mel-frequency cepstral coefficient (MFCC) and Gaussian mixture model (GMM) classifier. MFCCs were calculated using 40 ms frames with a Hamming window, 50% overlap and 40 mel bands. For classification, the first 20 coefficients were kept, including the 0th order coefficient, along with delta and acceleration coefficients calculated using a window length of 9 frames. The 0th order MFCC was included in the feature vector for keeping information on the energy of the signal, which may provide discriminative information for certain scene classes. Each acoustic scene was modeled using a 16-component GMM trained using the expectation maximization algorithm. During testing, predictions were obtained using maximum likelihood classification among all available models, with likelihood accumulated over the entire test signal.

Classification performance is measured using accuracy, representing the number of correctly classified segments among the total number of test segments. The overall classification accuracy of the baseline system on the development data, obtained using the provided cross-validation setup, is 72.5%, with class-wise performance ranging from 13.9% to 98.6%. The baseline system classification accuracy on the subsequently released evaluation set is 77.2%. The baseline system is marked in the results as DCASE.

C. Challenge results

As seen in Table II, Task 1 is the most popular task of the 2016 challenge, with a total of 48 submissions from 34 different teams. Most submitted systems outperform the baseline system, which is expected, given its simplicity. Out of 48 submitted systems, 22 use deep learning (DL), and 7 teams use the binaural input or multiple combinations of the two audio channels.

Various classification approaches were used, including feed-forward neural networks, recurrent (RNN, including LSTM), convolutional (CNN), and combinations of neural networks with other techniques, specifically GMMs. SVM-based approaches account for 10 submitted systems, while ensemble classifiers are used in 10 other systems. The list of top performers is dominated by ensemble classifiers [41]–[43] and deep learning classification methods, in particular CNNs [12], [29], [44]. We also note that factor analysis methods perform well: i-vectors [41] and NMF [45] are among top performing systems, exploiting the fact that each scene is composed of multiple sources whose joint variations can be explained using latent variables. Table III summarises top-performing systems, including information on the features and

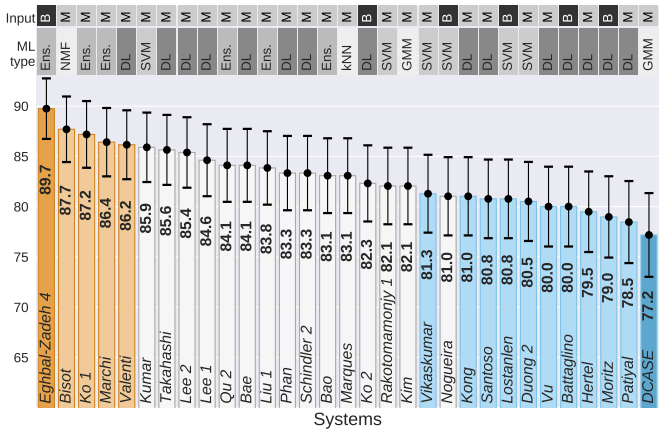

Fig. 3. Acoustic scene classification task accuracies based on the evaluation set with 95% confidence intervals, selected top system per participant. Based on McNemar's test with a significance level of 0.05, 4 runners-up systems cannot be judged to perform differently to the winner (marked in orange), and a number of systems do not perform differently to the baseline system (blue) under the same statistical test conditions.

TABLE III
SELECTED TOP RANKED SYSTEMS SUBMITTED FOR ACOUSTIC SCENE CLASSIFICATION TASK.

System	Features	Classifier	Acc
Eghbal-Zadeh 4	MFCC+spectrogram	ensemble	89.7 %
Bisot	spectrogram	NMF	87.7 %
Ko 1	various features	ensemble	87.2 %
Marchi	various features	ensemble	86.4 %
Valenti	mel energies	CNN	86.2 %
Kumar	MFCC distribution	SVM	85.9 %
Takahashi	MFCC	DNN-GMM	85.6 %
Lee 2	unsupervised features	CNN ensemble	85.4 %
Bae	spectrogram	CNN-RNN	84.6 %

classification approach employed in each system. Figure 3 lists those systems that outperform the baseline, including details on use of monophonic (M) or binaural (B) audio, and machine learning approach.

From a feature design perspective, representations using the mel-frequency scale (MFCCs and log-mel energies) were most popular among the 48 submissions. The main reason for this is that they provide a reasonably good representation of the spectral properties of the signal and provide reasonably high inter-class variability to allow class discrimination by many different machine learning approaches. Other choices included CQT-based time-frequency representations [45], combinations of various features (including mel-based) [41]–[43], and representations learned in an unsupervised way [46].

D. Discussion

Even though many of the top performing systems were based on deep learning methods, the evaluation shows that good performance can be obtained using classical methods too, such as SVM or NMF. Comparing performance between the development and evaluation datasets, most systems have similar or better performance on the evaluation set, showing that they exhibit good generalization properties.

Confidence intervals, calculated as a binomial proportion confidence interval for the classification output being correct or incorrect with respect to the ground truth, are presented in Fig. 3 for selected top systems per participant. It can be seen that the confidence intervals of systems with similar performance overlap significantly. A further analysis of the classification output using McNemar's test for comparing classifiers [47] shows that some systems cannot be considered as performing differently than the winner for a significance level of 0.05. Similarly, a number of systems cannot be differentiated from the baseline system under same statistical test conditions. Class-wise results show rather large difference in classification performance between systems and for different scene categories, with most difficult classes being library (lowest score obtained by at least one system 0%) and train (11.5%), while beach, bus, car and office had a score of at least 69% for all systems.

A listening experiment based on the evaluation dataset was set up for comparing systems' performance to human performance. Due to the size of the dataset, subsets containing 30 audio segments were presented to each test subject, with two segments for each scene class. The test segments per subject were randomly selected without replacement, resulting in the complete evaluation dataset being distributed among 13 test subjects.

A total of 87 participants provided 2 610 individual task answers. For evaluation, each audio sample is considered a separate test item and compared to the corresponding ground truth. The overall performance of the human subjects calculated over all answers was 54.4 %, while average performance across contexts for all submitted systems is 80.9% - the difference in performance is striking. Previous similar experiments resulted in human performance similar or higher than that of automatic classification methods using same data: for example in [8], human performance was 69% for 24 classes and 88% for 6 classes, just slightly higher than the automatic methods proposed in the same work; human performance for the 10 classes of DCASE 2013 data was determined to be 72% [7] and 79% [48] in two different setups, both being much better than the 55% average of the submissions. A breakdown of subjects into groups shows that the ones familiar with the Finnish soundscape had an average recognition accuracy of 60% compared to the participants from outside Finland that reached only 53%. At the same time, an expert listener who was highly familiar with the data and tested with the entire evaluation set obtained a performance of 77%.

Confusion matrices for the submitted systems and human ratings are presented in Figures 4 and 5. Some similarities can be observed, for example in the confusion of park and residential area, and train being confused with cafeteria in recordings made in the train's restaurant car. Other confusions are understandable from a human perspective, such as not distinguishing easily between forest path and park or city center and residential area streets, while another notable confusion of automatic systems is between home and library. The poor performance of humans is rather surprising, but could be explained by lack of familiarity of the subjects with the acoustic characteristics of the scene, and the small amount

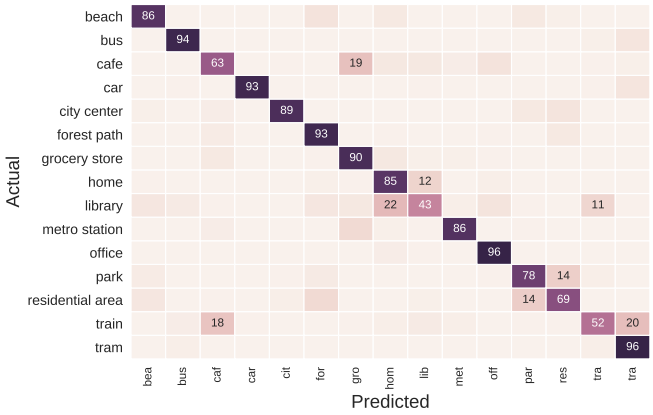

Fig. 4. Confusion matrix for all submitted systems

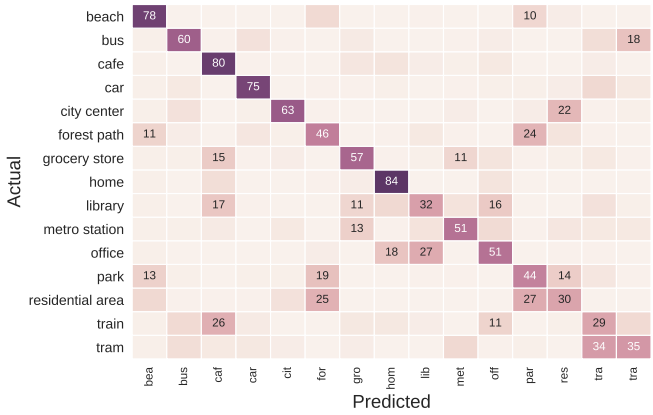

Fig. 5. Confusion matrix for human classification

of training data offered in the familiarization stage of the listening experiment. In addition, test subjects were allowed to answer whenever ready, thus reducing the amount of acoustic information used in the decision making process. A closer investigation of the listening test results is presented in [49].

IV. Sound event detection in synthetic audio

The goal of Task 2 is to detect possibly overlapping sound events, using synthetic mixtures simulating an office environment. As such, the task is directly related to the problem of *polyphonic sound event detection* and is a successor of the Event Detection - Office Synthetic task carried out at DCASE 2013 [38]. By using synthetic mixtures, Task 2 studies the behavior of tested algorithms when facing controlled levels of complexity (noise, polyphony), with the added benefit of a very accurate ground truth.

A. Dataset and experimental setup

Audio data for Task 2 contains instances of 11 sound classes related to office sounds: clearing throat, coughing, door knock, door slam, drawer, human laughter, keyboard, keys (placed on a table), page turning, phone ringing, and speech. Audio sequences for this task were created from isolated sound events using the sound scene synthesizer of [50]. Recordings of isolated sound events were made at LS2N, École Centrale de Nantes, using a shotgun microphone AT8035 connected to a ZOOM H4n recorder. Audio files are sampled at 44.1 kHz and are monophonic.

The task involves three datasets: training, development, and testing. The training set contains recordings of 20 isolated sounds per class, for the 11 classes enlisted above. The development dataset contains 18 simulated sound scenes of 2 min duration each generated using the same isolated segments found in the training dataset, plus background sounds. Finally, the test dataset contains 54 audio files of simulated sound scenes of 2 min duration each, using a pool of 440 isolated event segments not available in the training and development datasets, plus background sounds also different from the one used in the development dataset. The development and test datasets contain ground truth annotations automatically generated by the sound scene synthesizer, in the form of a sound event list identified by a start time, end time, and sound event class.

Parameters controlling the simulated material include the event-to-background ratio (EBR), the presence/absence of overlapping events (monophonic/polyphonic scene), and the number of active events per class. The EBR of an event of length N (in samples) is obtained by computing the ratio in decibel between the event E_{rms} and the background B_{rms} root mean square measures:

$$EBR = 20\log_{10}\left(\frac{E_{rms}}{B_{rms}}\right) \tag{1}$$

where E_{rms} and B_{rms} are defined as:

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

with $x(n)$ being replaced by either $e(n)$ or $b(n)$, the sound pressures at sample n of respectively the sound event sequence and the background noise. In the Task 2 dataset, the EBR has values of -6, 0, and 6 dB. For monophonic scenes, the number of active events per class varies from 1 to 3 and for polyphonic scenes from 3 to 5.

B. Baseline system and evaluation metrics

The baseline system developed for this task is based on supervised non-negative matrix factorization (NMF) [51] and uses a dictionary of pre-extracted spectral templates, created using the training dataset. For pre-processing, the system computes a variable-Q transform (VQT) spectrogram [52] with a 10 ms time step and a log-frequency resolution of 60 bins/octave. A simple noise removal process detects silent regions in the recording and uses them as the noise level. Supervised NMF with beta-divergence and 30 iterations is used to decompose the VQT spectrogram into a pre-extracted and fixed spectral basis matrix (estimated during training) and a sound event activation matrix. The latter matrix is subsequently thresholded and post-processed into a list of detected events per time frame.

Following a community discussion using the DCASE 2016 mailing list, a set of evaluation metrics for sound event detection was chosen. The metrics are presented in detail

TABLE IV
SUMMARY OF SYSTEMS SUBMITTED FOR THE SOUND EVENT DETECTION IN SYNTHETIC AUDIO TASK.

System	Features	Classifier	ER_{1s}	F_{1s}
Komatsu	variable Q transform	semi-supervised NMF	0.33	80.2 %
Choi	mel energy	DNN	0.36	78.7 %
Hayashi	mel filterbank	BLSTM	0.40	78.1 %
Phan	Gammatone cepstrum	Random forests	0.59	64.8 %
Giannoulis	various	CNMF	0.67	55.8 %
Pikrakis	Bark scale coefficients	Template matching	0.74	37.4 %
Vu	CQT	RNN	0.89	52.8 %
Gutierrez	MFCC	kNN	2.08	25.0 %
Kong	mel filterbank	DNN	3.54	12.6 %

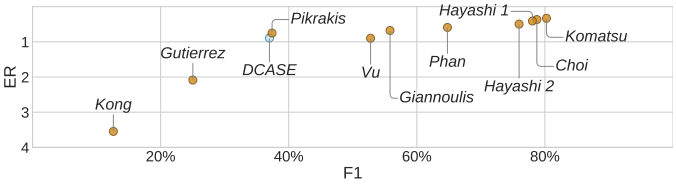

Fig. 6. Task 2 results on the evaluation set: segment-based error rate vs. F-score for all submitted systems. The baseline system is marked in blue and is ranked 8th of 11 systems.

in [53]. In Task 2, the main metric is the segment-based total error rate evaluated in one second segments over the entire test set, denoted ER_{1s}. In segment-based metrics, an event in the system output is considered correctly detected if its temporal position overlaps with the segment of an event with the same label in the ground truth. Additional metrics for Task 2 include the segment-based F-score, denoted as F_{1s}, and the onset-only event-based F-score with 200ms tolerance. Performance of the baseline system for the development dataset is $ER_{1s} = 0.78$, whereas for the test dataset $ER_{1s} = 0.89$.

C. Challenge results

Task 2 had 10 submissions from 9 teams, as can be seen in Table II. In terms of the error rate, 6 submissions outperformed the baseline system. Results in terms of the segment-based error rate and F-measure are shown in Table IV, with a graphical representation of the results shown in Fig. 6.

As can be seen from Table IV, about half of the submissions use some form of deep learning, including feedforward networks, recurrent neural networks (RNNs), bi-directional long short memory networks (BLSTMs), and convolutional neural networks (CNNs). The BLSTMs were also combined with hidden Markov models (HMMs) for modelling sound event durations. Two submissions are based on non-negative matrix factorisation (NMF), one using convolutive NMF. There were also approaches that used random forests, k-nearest neighbors and template matching. For both sets of metrics, the best performing system used NMF with a mixture of local dictionaries, combined with SVM postprocessing.

In terms of features, most approaches used time-frequency representations, including the constant-Q transform (CQT), variable-Q transform (VQT), and mel spectrograms. When compared with Task 1, the extracted features have a higher frequency resolution, in order to disambiguate multiple overlapping sound events. Other features used included MFCCs, gammatone cepstra, and Bark scale coefficients.

D. Discussion

A few submitted systems reported $ER_{1s} > 1$, which indicates that the F-measure was used as the main metric for training the submitted systems. Still, there is an almost perfect agreement with respect to rankings when comparing the error rate with the F-measure. Segment-based scores are generally higher compared to event-based scores (even considering that event-based scores only consider the sound event onset and not the offset). This drop for event-based metrics implies the lack of either temporal precision or temporal tracking in submitted systems.

With respect to the generalization capabilities of the systems, most report a significant drop in performance ($10-30\%$ in terms of absolute F-measure) when compared with development set results. This was mostly observed in conjunction with neural network-based systems, which might imply overfitting, most probably because the development set used the same samples as the training set. As expected, results depend on the sound class. For example, the first-ranked system of Komatsu et al. [54] reports an F-measure of 90.7% on door knock events, and a 37.7% on door slams. The door slam class in particular was the most challenging to detect amongst all systems, possibly due to the short duration of such events.

Due to the nature of the dataset, where groups of recordings have specific properties with respect to EBR and event density, an analysis of overall system performance for Task 2 is performed using a one-way repeated measures ANOVA. Sphericity is evaluated according to a Maulchy test [55], using a significance threshold of 0.05. This analysis, performed on the class-wise event-based F-measure, shows that out of 10 systems, 7 significantly improve upon the baseline system. This metric is chosen as it is not sensitive to the duration and density (number of events per scene) of the events.

When comparing the performance of systems to detect monophonic vs. polyphonic sequences, the ANOVA analysis does not indicate any significant difference. Results with respect to background noise show that the higher the EBR, the better is the performance of the systems (with the exception of the system of Komatsu et al [54]). Only four systems have significantly better performance than the baseline for all EBR levels. Finally, statistical significance evaluation w.r.t. the num-

ber of events does not show any influence of this parameter on the performance of the evaluated systems. A detailed statistical analysis of results for Task 2 is provided in [56].

V. Sound event detection in real-life audio

Task 3 evaluates the performance of sound event detection systems in multi-source conditions similar to everyday life, where sound sources are rarely heard in isolation. Contrary to the synthetic audio task, there is no control over the number of overlapping sound events at each time, both in the training and testing audio data.

A. Dataset and experimental setup

Task 3 uses the TUT Sound Events 2016 dataset, consisting of two common everyday environments: one outdoor (residential area) and one indoor (home). The audio material consists of the original full length recordings that are also part of TUT Acoustic Scenes with the same scene label. Target sound event classes were selected based on their frequency in the annotations. The annotations were produced by two research assistants, using nouns to characterize the sound source, and verbs to characterize the sound production mechanism, wherever this was possible. The full recording and annotation procedure is described in [40].

The event classes and the number of examples available for each class are listed in Table V. The recordings contain many other overlapping sounds, but only the listed classes are considered for the current detection task. Two sets of annotations were provided: the simplified annotations containing only the selected classes, and the full annotation containing all available annotated sounds, with the baseline system implementation based on the simplified annotation set.

The data were partitioned so that a higher proportion of instances for each class were included in the development set. This resulted in keeping 5 recordings for evaluation in each scene. The development set consists of 12 recordings for residential area having 60-80% of total available instances per class and 10 recordings for home, having 40-80% of instances. The provided cross-validation setup for the development set consists of 4 folds, in which each recording is tested exactly once.

TABLE V
TUT Sound Events 2016: Most frequent event classes with number of instances

Residential area		**Home**	
event class	instances	event class	instances
(object) banging	23	(object) rustling	60
bird singing	271	(object) snapping	57
car passing by	108	cupboard	40
children shouting	31	cutlery	76
people speaking	52	dishes	151
people walking	44	drawer	51
wind blowing	30	glass jingling	36
		object impact	250
		people walking	54
		washing dishes	84
		water tap running	47

B. Baseline system and evaluation metric

The baseline system provided for the task is based on the same method as used in Task 1. It uses MFCCs and GMMs, with MFCCs calculated using the same parameters as in the baseline system for Task 1. For each event class, a binary classifier is used, with the positive class model trained using those audio segments annotated as belonging to the modeled event class, and a negative class model trained using the remainder of the audio recording [40]. During testing, the decision for each event class is independent, based on computing the likelihood ratio between positive and negative models for the class within a one second sliding window.

Evaluation of system performance for sound event detection uses as the primary metric the segment-based error rate in one second segments, as in Task 2. Secondary metrics are the segment-based F-score and event-based error rate and F-score. The segment-based error rate of the baseline system on the development set is 0.91, while on the evaluation dataset it is 0.88. For the evaluation stage, the system was trained using the full development set, resulting in better performance due to availability of more training data.

C. Challenge results

There were 16 submissions for Task 3, originating from 12 different teams. Surprisingly, only one of the submitted systems performed better than the baseline system in terms of segment-based error rate. Systems based on deep learning accounted for most of the systems, with top 7 submissions based on DNN, RNN or fusion including deep learning architectures. Other classification approaches include random forests and one GMM-HMM solution. A system generating random events for each one second segment was also submitted, to simulate a data-driven solution tailored to the evaluation metric and using only statistics of the annotation, disregarding the audio completely. Unsurprisingly, it ranked very low.

The choice of features is dominated by mel representations: out of 16 systems, 9 use MFCCs and 4 use mel energies. The most obvious explanation for this is that MFCCs and mel energies provide a compact yet reasonably informative representation of the signal spectrum. Only one team (two submissions) exploited binaural acoustic information [57].

The segment-based performance of all submitted systems is presented in Figure 7, and top three systems according to ER are summarized in Table VI. The scatter plot in Figure 7 places the best system closest to the upper right corner. It can be noticed that 8 submissions had better F-score than the baseline system. The top system has $ER_{1s} = 0.80$, which is relatively high, considering that a zero-output system has $ER_{1s} = 1$ [53]. The F-score of the top system is however also the highest of all submissions, at 47.8 %. The runner-up in terms of ER is the baseline system, while for F-score two other submissions obtain 41.9 % and 41.1 %, respectively. Most submissions had error rates between 0.9 and 1.

D. Discussion

The trend for using deep learning is evident also for Task 3. The structure and training of neural networks allow directly

TABLE VI
SUMMARY OF SELECTED SYSTEMS SUBMITTED FOR SOUND EVENT DETECTION IN REAL LIFE AUDIO TASK.

System	Features	Classifier	ER_{1s}	F_{1s}
Adavanne 1	mel energy	RNN	0.80	47.8 %
Zoehrer	spectrogram	GRNN	0.90	39.6 %
Vu	mel energy	RNN	0.91	41.9 %
Liu	MFCC	fusion	0.92	34.5 %
Kong	MFCC	DNN	0.95	36.3 %
Pham	MFCC	DNN	0.95	11.6 %
Elizalde4	MFCC	Random Forests	0.96	33.6 %
Phan	GCC	Random Forests	0.96	23.9 %
Gorin	mel energy	CNN	0.97	41.1 %

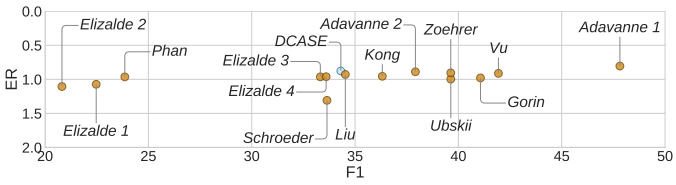

Fig. 7. Task 3 (Sound event detection in real life recordings) results using evaluation data: segment-based error rate and F-score for all submitted systems. The baseline system is marked in blue and has the second smallest error rate.

and very easily a setup for multi-label classification, which fits the task of polyphonic sound event detection. On the other hand, due to some classes having a small number of instances, most methods, and especially the deep learning methods, fail to detect them, being optimized to detect most of the events belonging to more frequent classes. A look at class-wise performance reveals that the top system detects only few classes, with F-scores 76 % for water tap and 16.5 % for washing dishes in home scenes, 62 % for bird singing, 76.7 % for car passing by, 32 % for wind blowing in residential area scenes, and all other classes 0 %. Full class-wise results are available on the challenge webpage [58].

A close look at the scene-wise performance reveals that sound events in residential area scenes were easier to detect ($ER_{1s} = 0.78$) than the ones in home scenes ($ER_{1s} = 0.91$). This is likely due to residential area classes being clearly distinct, while home classes are more similar to each other: in residential area scenes, the sound event classes are mostly related to concrete physical sound sources (bird singing, car passing by), while the home scenes are dominated by abstract object impact sounds (dishes, cutlery, etc).

Tasks 2 and 3 address the same problem and use the same metrics, but use different material (synthetic vs. real audio), resulting in a large difference in results: error rate 0.33 and F-score 80.2 % top score for synthetic data, while for real audio top scores are 0.81 and 47.8 %, respectively. This difference can be explained by the complexity of the audio: Task 2 synthetic data were generated with a controlled number of overlapping events and a quiet background, while Task 3 data have an unknown number of overlapping events, including sounds not belonging the target classes. Part of the difference in systems' performance can also stem from the manual annotation of real-world data, as manual annotations are inherently noisy and this affects both evaluation scores and training methods. Results achieved for Task 3 demonstrate the difficulty of the event detection task in a realistic setting.

VI. DOMESTIC AUDIO TAGGING

Task 4 is based on audio recordings made in a domestic environment. It involves multi-label classification of 4-second audio chunks, with the set of label classes based on prominent sound sources in the acoustic environment. For a given audio chunk, submitted systems are required to output a classification score for each of the seven label classes listed in Table VII. In the available development dataset, multi-label annotations are provided for each audio chunk.

A. Dataset and experimental setup

The audio recordings used in Task 4 originate from the Computational Hearing in Multisource Environments (CHiME) project [59], [60]. These recordings were subsequently annotated and released as CHiME-Home [32], a multi-annotation dataset aimed at audio tagging tasks.

1) Audio recordings: The CHiME-Home dataset consists of approx. 6.8 hours of stereophonic audio, obtained by positioning binaural recording equipment inside a house. The acoustic environment comprises the following sound sources: Two adults and two children, television and electronic gadgets, kitchen appliances, footsteps and knocks produced by human activity, further to sound originating from outside the house.

In Task 4, audio data are provided at sampling rates 48 kHz and 16 kHz, respectively as stereophonic and monophonic recordings. The 16 kHz recordings were obtained by down-sampling the right-hand channel of the 48 kHz recordings. All audio data are available for system development, however the subsequent evaluation is performed using the monophonic audio sampled at 16 kHz. This approach aims at approximating the recording capabilities of commodity hardware.

2) Annotations: The audio was partitioned into 6 137 non-overlapping 4-second audio chunks. Subsequently, three human annotators were each asked to assign labels to each of the chunks. The set of possible label classes included those listed in Table VII, with two auxiliary labels for flagging chunks as silent or unidentifiable. To increase confidence about annotations, Task 4 is evaluated using only the chunks for which two or more annotators assigned the same label, for all considered labels. The final labels of those 2 762 chunks with 'strong agreement' between annotators are then determined by majority voting across annotators.

3) Development and evaluation data: Out of 6 137 chunks, 4 378 chunks are available for system development, with the remaining 1 759 chunks previously reserved for release after DCASE 2016, by partitioning at the level of 5-minute recording segments. The 4 378 chunks in the development dataset include 1 946 'strong agreement' chunks for training and testing. The remaining 2 432 chunks in the development dataset are available as additional training material. Out of the 1 759 chunks reserved for release after conclusion of DCASE 2016, there are 816 'strong agreement' chunks. We use these 816 chunks as evaluation data. Table VII reports label occurrences for 'strong agreement' chunks. To help quantify

TABLE VII
LABEL OCCURRENCES IN DEVELOPMENT AND EVALUATION SUBSETS OF DOMESTIC AUDIO TAGGING TASK. BASED ON AUDIO CHUNKS WITH STRONG ANNOTATOR AGREEMENT, COUNTS REPORTED IN BOLDFACE AND COUNTS WHERE ALL 3 ANNOTATORS AGREED IN ITALICS.

Label	Description	Number of audio chunks			
		Development		Evaluation	
c	Child speech	**1214**	*1143*	**328**	*301*
m	Adult male speech	**174**	*152*	**79**	*69*
f	Adult female speech	**409**	*339*	**140**	*126*
v	Video game/TV	**1181**	*1141*	**590**	*571*
p	Percussive sounds, e.g. crash, bang, knock, footsteps	**765**	*344*	**269**	*119*
b	Broadband noise, e.g. household appliances	**19**	*9*	**31**	*31*
o	Other identifiable sounds	**361**	*21*	**125**	*10*

annotator agreement, the table furthermore reports label occurrences where for a given label all 3 annotators agreed about its presence. For a discussion of annotator agreement, please refer to [32].

To aid system development, we further partition the development data at the level of 5-minute recording segments for 5-fold cross validation. In the partition, due to a low number of associated label occurrences, we omit the 5-minute recording constraint for chunks labelled 'b'.

B. Baseline system and evaluation metric

The baseline system for Task 4 relies on MFCCs combined with GMMs. For simplicity, the system is based on the same software implementation as Task 1 and Task 3. The chosen system parameters for Task 4 closely match those previously reported for the CHiME-Home dataset [32], parameters which we observed yielded favorable results. Thus, we obtain 20 ms frames with a Hamming window and 50% overlap. Subsequently, excluding the 0th order coefficient we extract the 13 first MFCCs, based on 40 mel frequency bands. Finally, after normalizing feature vectors to zero mean and unit variance, for each of the seven considered labels we train an independent binary classifier consisting of two 8-component GMMs. Given a set of input frames, the label-wise classification score is the log-likelihood ratio of the two associated GMMs.

To quantify prediction performance with respect to a given label, we follow the convention of considering a range of possible classifier operating points. We compute the equal error rate (EER) [61], which is the fixed point of the graph of false negative rate versus false positive rate, plotted in response to the operating point. Thus, the EER approximates the classification error rate we would obtain for equal amounts of positive and negative instances, facilitating comparison of performance across label classes. Averaged across labels, the baseline system yields EERs 0.213 and 0.209, for development and evaluation datasets, respectively.

C. Challenge results

With eight submissions by seven teams, Table VIII displays obtained EERs for each of the seven individual labels, in addition to label-averaged EERs used to rank the submissions. As observed, in terms of label-averaged EERs, with the exception of two systems all submissions outperform the baseline. Obtained label-averaged EERs range from 0.166 to 0.221, with the best-performing and worst-performing submissions respectively representing a performance gain of 20.6% and a performance loss of 5.7%, relative to the baseline.

As was observed across all DCASE 2016 tasks, neural networks are a popular choice of classification technique, comprising seven out of eight submissions for Task 4. The two best-performing submissions rely on convolutional architectures. These results notwithstanding, we observe that the submission ranked third outperforms the baseline by 16.7%, while still based on GMMs. All submissions rely on widely-applied input features, with the top three submissions based on CQT features, mel spectrograms and MFCCs respectively. Across submissions, the most popular features are MFCCs.

D. Discussion

Examining label-wise EERs, averaged across submissions, we observe that the two least challenging labels are v and b, with respective mean EERs 0.061 and 0.084. Analogously averaging across submissions, the two most challenging labels are m and o, with respective mean EERs 0.267 and 0.271. The remaining labels c, p, f have the associated mean EERs 0.205, 0.218, 0.241, respectively.

As previously noted [32], a possible explanation for such variation in submission-averaged performance across labels is that perceptually salient acoustic events are relatively easy to identify: Firstly, we expect the chosen audio features to represent predominantly those events occurring in the acoustic foreground, as opposed to those events in the acoustic background. Secondly, we expect those events which occupy relatively long segments within the 4-second chunks to be readily identifiable, due to relative abundance of relevant frames for training models and building predictions. Our own informal listening suggests that sources associated with labels v and b indeed are relatively perceptually salient, frequently occupying the entire duration of audio chunks. By comparison, human utterances (labels c, m, f) are shorter in duration. Nonetheless, among human speakers, child speech appears to strongly occupy the acoustic foreground.

Table VIII indicates that the submission rankings that we obtain with respect to individual labels may deviate from the label-averaged ranking. To quantify such discrepancy between rankings, for each label we compute Spearman's ρ between the EERs obtained for the given label, and the label-averaged EERs. Notably, we observe negative rank correlations for labels o and c, with ρ respectively -0.36 and -0.30. A possible explanation for the observed behaviour is that relevant acoustic events in chunks labelled o and c have relatively large acoustic variability are hence more prone to overfitting: For labels with large acoustic variability, we expect the relevant structure in the data to be less discernable, owing to relative data scarcity. This explanation appears consistent with the observation that the GMM-based approach submitted by Yun et al. [62] outperforms the ANN-based approaches submitted by Lidy et al. [63] and Cakir et al. [20] for label c. That the latter two submissions yield superior performance for labels m and v further suggests an advantage of ANNs combined with time-frequency input features compared to approaches based

TABLE VIII
DOMESTIC AUDIO TAGGING TASK RESULTS FOR EVALUATION DATASET, QUANTIFIED USING EQUAL ERROR RATE (EER) AND RANKED BY EER AVERAGED ACROSS LABELS.

System	Features	Classifier	Label-wise EER							Mean EER
				c	m	f	v	p	b	o
Lidy	CQT features	CNN	0.210	0.182	0.214	0.035	**0.168**	0.032	0.320	**0.166**
Cakir	Mel spectrogram	CNN	0.250	**0.159**	0.250	**0.027**	0.208	**0.022**	0.258	0.168
Yun	MFCCs	GMM	**0.177**	0.253	**0.179**	0.102	0.207	0.032	0.266	0.174
Kong	Mel spectrogram	DNN	0.195	0.280	0.229	0.090	0.221	0.039	0.272	0.189
Xu1	MFCCs	DNN	0.209	0.313	0.216	0.040	0.249	0.065	0.272	0.195
Xu2	MFCCs	DNN	0.203	0.304	0.236	0.037	0.275	0.048	0.280	0.198
DCASE	MFCCs	GMM	0.191	0.326	0.314	0.056	0.212	0.117	0.249	0.209
Vu	MFCCs	RNN	0.226	0.307	0.293	0.078	0.218	0.078	0.279	0.211
Hertel[a]	Magnitude spectrogram	CNN	0.183	0.278	0.234	0.080	0.201	0.323	**0.246**	0.221

[a]The submission by L. Hertel et al. was re-submitted after the deadline. The revised submission yielded substantially lower EERs, with the difference in performance attributed to a software bug in the original submission.

on MFCCs, for representing and identifying events occurring in the acoustic background.

To determine statistical significance of differences in performance, for each label and for each pair of submissions we apply the sign test [64] to bootstrapped paired samples of EERs. Observing that bootstrapped samples of EERs are not guaranteed to be symmetric about the median for label b, we motivate use of the sign test based on its few assumptions about underlying distributions. With the exception of the submission pair 'Vu' and 'Xu1' for label m, we observe that $p \ll 0.001$ for all combinations of submissions and labels.

VII. DISCUSSION

At first glance, deep learning methods stand out as the most employed approach among submitted systems. The emergence of neural network based methods is also obvious in the comparison with DCASE 2013, where there were no systems involving DNNs. It is likely that besides the general popularity of deep learning as a novel technique, the amount of data available in DCASE 2016 encouraged, and to a certain extent supported their use. However, at least for the sound event detection tasks, the data size was still insufficiently large to allow robust learning. In parallel with neural networks dominating algorithm choice, it appears that data-driven approaches tend to replace manual design. The combination of these factors calls for more data, and this was seen by the participants as the main aspect that needs improvement.

The acoustic scene classification task represents the most straightforward supervised classification setup. For this reason, Task 1 attracts interest through its possible uses in applications, as well as simplicity of deploying the familiar machine learning techniques that do not require significant modifications for this task. The latter is likely the reason for which Task 1 had the highest number of participants, serving as a very good entry level task for researchers starting work in the research field. The amount of data provided for the task allowed use of deep learning algorithms involving convolutive or recurrent networks.

Sound event detection (Tasks 2 and 3) represented a more difficult setup, and this resulted into a smaller number of participants trying to tackle the problem. Participants' opinions gathered using a survey after the challenge indicate dissatisfaction with the data amount for both tasks and class balance in case of Task 3.

For Task 2 specifically, the use of simulated recordings is counterbalanced by data generated under various conditions with the benefit of a very accurate ground truth, which allowed a detailed analysis of system performance in terms of specific aspects (noise, event density, polyphony) that would hardly be possible when considering real-world data. Although we acknowledge that the sole use of simulated data cannot be considered for definitive ranking of systems, we believe that the described task design is of great interest when paired with evaluation on real world data. Despite technical improvements of the acoustic realism such as the use of reverberation filters and 3-dimensional positioning of the sources, one interesting avenue for improvements would be to design a task roughly following the evaluation procedure presented in [50]. There, systems are first evaluated against real-world data. Secondly, a synthetic dataset is designed mimicking the real-world data, ensuring that systems perform similarly. Lastly, the systems are evaluated against variants of the synthetic data. This procedure provides more grounding to the evaluation on synthetic data and can provide insights about the performance of systems for real-world data.

Naturally one should be careful in drawing conclusions obtained with simulated data only, since it is unlikely to present all the diversity present in real-world data. There are also some caveats in the use of synthetic data, that can very easily lead to erroneous conclusions. For example, one should be very careful when combining samples from multiple sample databases, since each database may have different characteristics such as audio quality, recording device, etc. In such a case, instead of recognizing target sound classes, an analysis system may learn to recognize these database-specific factors. Obviously, the same audio sample should not be present in the dataset multiple times, and definitely not in both the training and testing sets, in order to prevent overfitting. Ideally, some synthesis process could be used to produce large quantities of training material, but testing would be done with smaller amount of carefully annotated real material.

The limited amount of data available for some classes in Task 3 resulted in them not being detected at all by some systems. This indicates that the optimization process was guided by the activity of larger classes, in detriment of the

classes that were insufficiently represented. However, such data are a truthful representation of real world situations, therefore in order to detect the less common sound events, systems should deal with data imbalance rather than requiring better balanced datasets.

Among the four tasks, a unique aspect of the audio tagging task is its reliance on monophonic, downsampled audio as a means of simulating the recording capabilities of commodity hardware. To better establish the effect of such audio degradation on performance, future investigations should quantify label prediction accuracy in response to a range of downsampling factors and simulated microphone characteristics.

With respect to evaluation metrics, it is worth pointing out that the current definition of segment-based metrics might not necessarily be the best way of measuring performance for sound event detection. A segment-based metric considers an event detected within a segment even if the sound is marked active for a very short duration within the segment. An event detected 10 ms early w.r.t. to reference annotation is considered correct in terms of event-based metrics, but may cause a false alarm if the 10 ms tolerance falls within the preceding segment in the segment-based metric. As a future recommendation, a rule on when an event should be considered active within a segment could be implemented, e.g. if the event is active at least 50% of the segment duration. A specific issue with the error rate is that it can result in scores surpassing 1, which can lead to interpretability issues; for this reason it is useful to also compute the F-score, to ensure that the measured output is plausible, even though it may contain many detection errors.

The area of sound scene analysis is increasingly active and there appears to be a need to maintain public evaluation campaigns such as DCASE in the foreseeable future, not only in order to evaluate the performance of state-of-the-art systems but also to serve as a focal point for the emerging research community. At the same time, issues around the long-term sustainability of the challenge need to be addressed. Based on the past challenges, a series of observations can be made:

- *Challenge organization*: Regarding central challenge organization, while DCASE 2013 and 2016 were efforts initiated by specific institutions and research groups, a direct outcome following the completion of DCASE 2016 was the establishment of a steering group comprising academics and researchers across several academic institutions and the industry, in an effort to provide high-level advice on the challenge organization. This was done in conjunction with receiving feedback from the IEEE AASP Technical Committee on Audio and Acoustic Signal Processing, as part of the committee's Challenges subgroup. A possible future direction, inspired by the MIREX challenge on music information retrieval [39] would be to de-centralize the organization of the various challenge tasks. This would enable the involvement of additional research groups and would also provide towards the long-term sustainability of the challenge by not putting too much organizational effort towards a specific research group.
- *Data collection*: Given the current setup of DCASE 2016 and the upcoming 2017 edition, where participants are given access to unlabeled test data, there is a need to produce new datasets for each new version of the challenge. Currently the reference annotations of the evaluation dataset are published after the challenge concludes. This creates sustainability issues, which can be addressed by creating artificial datasets (for example using the sound scene synthesizer of [50], as was done for DCASE 2013 and 2016), or by relaxing challenge assumptions by allowing reuse and extension of past challenge datasets.
- *Introducing new tasks*: DCASE 2013 included two tasks on sound scene classification and sound event detection, whereas the 2016 version additionally included an audio tagging task. As the field evolves, there is a need to introduce new tasks on other areas of sound scene analysis. For DCASE 2017 new tasks were introduced in the topics of rare sound event detection and weakly supervised sound event detection. As with the 'challenge organization' point above, the DCASE steering committee and its community mailing list can serve as a first point of contact towards introducing new tasks, or developing the previous tasks to make them more realistic and useful for the community.
- *Evaluation metrics*: As observed in community discussions and from results of submitted systems, there is not always an agreement on the use of evaluation metrics, which can be attributed to both disciplinary practices as well as a focus on specific applications. As part of future challenges, we note the importance of incorporating new evaluation metrics through community discussion in the DCASE mailing list, whilst however always maintaining past metrics for compatibility and completeness purposes. This was achieved as part of DCASE 2016, where the metrics toolbox for sound event detection [53] incorporated all metrics defined for DCASE 2013. In the long term, such an effort could lead to a community-led sound scene analysis evaluation toolbox similar to the 'mir_eval' toolbox for music informatics[1]. To support development and testing of new metrics, and to allow measuring performance of DCASE 2016 submissions using other metrics than the ones provided in the challenge results, the system outputs of all submitted systems were also published [2] and are available for comparison against the reference annotations.
- *Baseline systems*: A major difference between DCASE 2013 and 2016 was the introduction of a unified baseline system for Tasks 1, 3 and 4, which maintained the same back-end processing and learning methods across all tasks. Having this unified approach for all tasks can be useful to make it easier to participate in multiple tasks, and to more easily transfer findings between tasks. On the other hand, specific tasks may require substantially different techniques, which may require baselines using different learning methods. Once the research field

[1]https://github.com/craffel/mir_eval
[2]https://doi.org/10.5281/zenodo.926660

matures, we recommend that the baseline system is advanced enough, so that surpassing the baseline system performance is likely to require submitted systems to incorporate novel techniques.

VIII. Conclusions

DCASE 2016 Challenge evaluated computational methods for analysis acoustic scenes and events. Publicly available datasets, common metrics and evaluation procedures, and publicly available baseline tools allowed evaluating different algorithms independently from applications they have been developed for. The challenge was a success in terms of participation, the high number of participants showing that the topics and proposed tasks are of great importance in current audio research, and in particular on the emerging area of computational sound scene analysis. The selected tasks represent a good characterization of current interest, from the more general acoustic scene classification and audio tagging topics, to the detailed temporal detection of individual sound events.

For upcoming challenges and workshops on the topic, it is important to follow the suggestions and interest of the scientific community in the process of tasks selection, and to get involved with industrial researchers in order to have a more complete view of the research field. This will allow the community to suggest and coordinate tasks for future challenges. With the help of a steering committee comprising domain experts, the proposed tasks will be evaluated for selecting the most interesting ones and for providing feedback on their setup.

References

[1] D. S. Pallett, "A look at NIST's benchmark ASR tests: past, present, and future," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2003. ASRU'03*. IEEE, 2003, pp. 483–488.

[2] J. S. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[3] G. Awad, C. G. M. Snoek, A. F. Smeaton, and G. Qunot, "TRECVid semantic indexing of video: A 6-year retrospective," *ITE Trans. on Media Technology and Applications*, vol. 4, no. 3, pp. 187–208, 2016, invited paper.

[4] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *12th International Conference on Latent Variable Analysis and Signal Separation*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds., 2015, pp. 387–395.

[5] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[6] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: A survey and a challenge," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept 2016, pp. 1–6.

[7] D. Barchiesi, D. Giannoulis, D. Stowell, and M. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.

[8] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, Jan 2006.

[9] 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," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[10] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio context recognition using audio event histograms," in *18th European Signal Processing Conference*, Aug 2010, pp. 1272–1276.

[11] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 1, pp. 142–153, Jan. 2015.

[12] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 95–99.

[13] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22st ACM International Conference on Multimedia (ACM-MM'14)*, Nov. 2014.

[14] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *18th European Signal Processing Conference (EUSIPCO 2010)*, 2010, pp. 1267–1271.

[15] S. Innami and H. Kasai, "NMF-based environmental sound source separation using time-variant gain features," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 1333 – 1342, 2012.

[16] J. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme, "An exemplar-based NMF approach to audio event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2013, pp. 1–4.

[17] A. Mesaros, O. Dikmen, T. Heittola, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 151–155.

[18] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 26, 2015.

[19] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[20] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. on Audio Speech and Language Processing*, 2017, arXiv preprint arXiv:1702.06286.

[21] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[22] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6440–6444.

[23] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green, "Automatic generation of social tags for music recommendation," in *Advances in Neural Information Processing Systems*, 2008, pp. 385–392.

[24] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[25] D. Tingle, Y. E. Kim, and D. Turnbull, "Exploring automatic music annotation with acoustically-objective tags," in *Proc. of the International Conference on Multimedia Information Retrieval*, 2010, pp. 55–62.

[26] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Sparse multi-label linear embedding nonnegative tensor factorization for automatic music tagging," in *18th European Signal Processing Conference*, 2010, pp. 492–496.

[27] E. Coviello, Y. Vaizman, A. B. Chan, and G. R. Lanckriet, "Multivariate autoregressive mixture models for music auto-tagging," in *International Conference on Music Information Retrieval*, 2012, pp. 547–552.

[28] K. Ellis, E. Coviello, A. B. Chan, and G. Lanckriet, "A bag of systems representation for music auto-tagging," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2554–2569, 2013.

[29] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," *arXiv preprint arXiv:1703.01789*, 2017.

[30] B. Defréville, F. Pachet, C. Rosin, and P. Roy, "Automatic recognition of urban sound sources," in *Audio Engineering Society Convention 120*, 2006.
[31] D. Stowell and M. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," in *Proc. of the AES 53rd International Conference: Semantic Audio*, 2014, pp. 80–86.
[32] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, "CHiME-Home: A dataset for sound source recognition in a domestic environment," in *Proc. of the 9th IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015.
[33] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proc. of the 2016 ACM on Multimedia Conference*, ser. MM '16. ACM, 2016, pp. 1038–1047.
[34] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, p. e488, 2014.
[35] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. Jackson, and M. D. Plumbley, "Unsupervised feature learning based on deep models for environmental audio tagging," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2017.
[36] T.-W. Su, J.-Y. Liu, and Y.-H. Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks."
[37] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
[38] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. on Multimedia*, vol. 17, no. 10, pp. 1733–1746, October 2015.
[39] "Music Information Retrieval Evaluation eXchange (MIREX)," http://music-ir.org/mirexwiki/.
[40] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, 2016.
[41] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks," http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification, DCASE2016 Challenge, Tech. Rep., September 2016, [Online; accessed 15-Mar-2017].
[42] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini, and B. Schuller, "Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 65–69.
[43] S. Park, S. Mun, Y. Lee, and H. Ko, "Score fusion of classification systems for acoustic scene classification," http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification, DCASE2016 Challenge, Tech. Rep., September 2016, [Online; accessed 15-Mar-2017].
[44] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 11–15.
[45] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification, DCASE2016 Challenge, Tech. Rep., September 2016, [Online; accessed 15-Mar-2017].
[46] J. Kim and K. Lee, "Empirical study on ensemble method of deep neural networks for acoustic scene classification," http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification, DCASE2016 Challenge, Tech. Rep., September 2016, [Online; accessed 15-Mar-2017].
[47] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
[48] J. Krijnders and G. t Holt, "Tone-fit and MFCC scene classification compared to human recognition," http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/abstracts/SC/KH.pdf, [Online; accessed 4-Apr-2017].
[49] A. Mesaros, T. Heittola, and T. Virtanen, "Assessment of human and machine performance in acoustic scene classification: DCASE 2016 case study," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, submitted.
[50] G. Lafay, M. Lagrange, M. Rossignol, E. Benetos, and A. Roebel, "A morphological model for simulating acoustic scenes and its application to sound event detection," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1854–1864, October 2016.
[51] D. D. Li and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.
[52] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *AES 53rd Conference on Semantic Audio*, January 2014, p. 8 pages.
[53] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016.
[54] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with mixtures of local dictionaries," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 45–49.
[55] R. Gueorguieva and J. Krystal, "Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry," *Archives of General Psychiatry*, vol. 61, no. 3, pp. 310–317, 2004.
[56] G. Lafay, E. Benetos, and M. Lagrange, "Sound event detection in synthetic audio: analysis of the DCASE 2016 task results," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2017.
[57] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 6–10.
[58] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. Plumbley, "Detection and classification of acoustic scenes and events DCASE2016," http://www.cs.tut.fi/sgn/arg/dcase2016/, 2016, [Online; accessed 4-Apr-2017].
[59] H. Christensen, J. Barker, N. Ma, and P. D. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. 11th INTERSPEECH Conf.*, 2010, pp. 1918–1921.
[60] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. D. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.
[61] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
[62] S. Yun, S. Kim, S. Moon, J. Cho, and T. Kim, "Discriminative training of GMM parameters for audio scene classification," DCASE2016 Challenge, Tech. Rep., September 2016.
[63] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification and domestic audio tagging," DCASE2016 Challenge, Tech. Rep., September 2016.
[64] W. J. Conover, *Practical Nonparametric Statistics*. John Wiley & Sons, 1980.

Annamaria Mesaros is a postdoctoral researcher at Laboratory of Signal Processing, Tampere University of Technology (TUT), Finland. She received the M.Sc. and Ph.D degrees in electronics and telecommunications in 2001 and 2007, respectively, from Technical University of Cluj Napoca, Romania, and Doctor of Science degree in signal processing from TUT in 2012. She has also been working as a postdoctoral researcher at Aalto University, Helsinki, Finland, within the Finnish Centre of Excellence in Computational Inference Research. Her research focuses on sound event detection in real-world multisource environments, including semantic aspects of human-generated sound annotation.

Toni Heittola received his M.Sc. degree in Information Technology from Tampere University of Technology (TUT), Finland, in 2004. He is currently pursuing the Ph.D. degree at TUT. His main research interests are sound event detection in real-life environments, sound scene classification and audio content analysis.

Emmanouil Benetos received the B.Sc. and M.Sc. degrees in informatics from the Aristotle University of Thessaloniki, Greece, in 2005 and 2007, respectively, and the Ph.D. degree in electronic engineering from Queen Mary University of London, U.K., in 2012. From 2013 to 2015, he was University Research Fellow with the Department of Computer Science, City, University of London, U.K. He is currently Lecturer and RAEng Research Fellow with the School of EECS, Queen Mary University of London, U.K. His research focuses on signal processing and machine learning for music and audio analysis, as well as applications to music information retrieval, acoustic scene analysis, and computational musicology.

Peter Foster is currently pursuing a career in industry with a focus on time series analysis. He received the Ph.D. degree from Queen Mary University of London, undertaken at the Centre for Digital Music, where he was subsequently employed as a postdoctoral research assistant. He received the M.Sc. and B.Sc. degrees in Computer Science from the University of Edinburgh and from the University of East Anglia, respectively. His interests are time series similarity and classification.

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

Tuomas Virtanen a professor at Laboratory of Signal Processing, Tampere University of Technology (TUT), Finland. He received the M.Sc. and Doctor of Science degrees in information technology from TUT in 2001 and 2006, respectively. He is known for his pioneering work on single-channel sound source separation using non-negative matrix factorization based techniques, and their application to noise-robust speech recognition, music content analysis and audio event detection. In addition to the above topics, his research interests include content analysis of audio signals in general and machine learning. He has authored about 100 scientific publications on the above topics. He is a member of the Audio and Acoustic Signal Processing Technical Committee of IEEE Signal Processing Society

Mark D. Plumbley (S'88-M'90-SM'12-F'15) received the B.A.(Hons.) degree in electrical sciences and the Ph.D. degree in neural networks from University of Cambridge, Cambridge, U.K., in 1984 and 1991, respectively. From 1991 to 2001, he was a Lecturer with Kings College London, London, U.K., before moving to Queen Mary University of London, London, in 2002, later becoming Director of the Centre for Digital Music. In 2015, he joined the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K., as Professor of Signal Processing. His research interests include automatic analysis of sounds and music, including acoustic scene analysis, audio source separation, and automatic music transcription, using methods such as deep learning, matrix factorization, and sparse representations. He is a Member of the IEEE Signal Processing Society Technical Committee on Signal Processing Theory and Methods.