TUT Database for Acoustic Scene Classification and Sound Event Detection

Annamaria Mesaros, Toni Heittola, Tuomas Virtanen
Department of Signal Processing
Tampere University of Technology
Tampere, Finland
email: annamaria.mesaros@tut.fi, toni.heittola@tut.fi,tuomas.virtanen@tut.fi

***Abstract*—We introduce TUT Acoustic Scenes 2016 database for environmental sound research, consisting of binaural recordings from 15 different acoustic environments. A subset of this database, called TUT Sound Events 2016, contains annotations for individual sound events, specifically created for sound event detection. TUT Sound Events 2016 consists of residential area and home environments, and is manually annotated to mark onset, offset and label of sound events. In this paper we present the recording and annotation procedure, the database content, a recommended cross-validation setup and performance of supervised acoustic scene classification system and event detection baseline system using mel frequency cepstral coefficients and Gaussian mixture models. The database is publicly released to provide support for algorithm development and common ground for comparison of different techniques.**

I. Introduction

Databases are of crucial importance in algorithm development, comparison of algorithms and reproducibility of results. Research fields that have well established benchmark databases benefit of rapid pace of development, with competition between teams on obtaining the highest performance. In this respect, detection and classification of acoustic scenes and events is picking up the pace, with special sessions organized in recent conferences and the Detection and Classification of Acoustic Scenes and Events (DCASE) 2013 challenge. This database is part of our effort to support interest in this research area and provide the research community with a starting point for data collection and common evaluation procedure.

Acoustic scene classification is defined as recognition of the audio environment, with applications in devices requiring environmental awareness [1], [2]. The environment can be defined based on physical or social context, e.g. park, office, meeting, etc. The problem is usually solved as a closed-set classification task, where identification of the current acoustic scene is required. A small number of publicly available datasets for acoustic scene classification exist. For example DCASE 2013 [3] acoustic scene development dataset contains 10 classes, 10 examples of 30 seconds length per class, with an evaluation set of the same size. Another example is the LITIS Rouen Audio scene dataset [4] containing 3026 examples for 19 classes, audio of length 30s. Additionally, a number of published studies use proprietary datasets. Results of acoustic scene classification range from 58% for 24 classes to 82% for 6 higher-level classes on the same data [2] to 93.4% for 19 classes [5]. Performance depends on the number of classes and their characteristics, with acoustic scenes that are very different from each other faring better, as expected.

Sound event detection is defined as recognition of individual sound events in audio, e.g. ”bird singing”, ”car passing by”, requiring estimation of onset and offset for distinct sound event instances and identification of the sound. Applications for sound event detection are found in surveillance, including security, healthcare and wildlife monitoring [6]–[12], audio and video content-based indexing and retrieval [13]–[15].

Sound event detection is usually approached as supervised learning, with sound event classes defined in advance and audio examples available for each class. Depending on the complexity of the required output, we differentiate between *monophonic sound event detection* in which the output is a sequence of the most prominent sound events at each time and *polyphonic sound event detection* in which detection of overlapping sounds is required [16]. Previous work on sound event detection is relatively fragmented, with studies using different, mostly proprietary datasets that are not openly available to other research groups. This hinders reproducibility and comparison of experiments. An effort in the direction of establishing a benchmark dataset was made with DCASE 2013 [3], by providing a public dataset and a challenge for different tasks in environmental sound classification. The training material contains 16 event classes, provided as isolated sound examples, 20 examples per class. The validation and evaluation data consist of synthetic mixtures containing overlapping events, 9 files for validation and 12 files for evaluation, with a length of over 1-2 minutes.

Collecting data for acoustic scene classification is a relatively quick process involving recording and annotation of audio. However, care should be taken to obtain high acoustic variability by recording in many different locations and situations for each scene class. On the other hand, annotation of audio recordings for sound event detection is a very slow process due to the presence of multiple overlapping sounds. An easier way to obtain well annotated data for sound event detection is creation of synthetic mixtures using isolated sound events - possibly allowing control of signal-to-noise ratio and amount of overlapping sounds [17]. This method has the

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

advantage of being efficient and providing a detailed and exact ground truth. However, synthetic mixtures cannot model the variability encountered in real life, where there is no control over the number and type of sound sources and their degree of overlapping. Real-life audio data is easy to collect, but is very time consuming to annotate.

We introduce a dataset of real-life recordings that offers high quality audio for research in acoustic scene classification and polyphonic sound event detection. The audio material was carefully recorded and annotated. A cross-validation setup is provided that places audio recorded in the same location to the same side of the experiment. This avoids contamination between train and test set through use of the exact same recording conditions, which can result in over-optimistic performance through learning of acoustic conditions instead of generalization.

The paper is organized as followes: Section II introduces the data collection principles, motivating the choices made in recording, annotation and postprocessing stages. Sections III and IV present in detail TUT Acoustic Scenes 2016 - the dataset for acoustic scene classification and TUT Sound Events 2016 - the dataset for sound event detection, including statistics about their content, partitioning for system development and evaluation, and performance of a simple baseline system in a cross-validation setup on the development set. The evaluation set was later released for the DCASE 2016 challenge [18]. Finally, Section V presents conclusions and future work.

II. Data collection principles

The data collection procedure takes into account the possibility for extending this dataset by other parties, therefore it includes some rules for recording and annotation to guarantee sufficient acoustic variability and uniform labeling procedure. The sound events dataset is planned as a subset of the acoustic scene dataset, by providing specific detailed annotations of sound event instances.

A. Recording

To satisfy the requirement for high acoustic variability for all acoustic scene categories, each recording was done in a different location: different streets, different parks, different homes. High quality binaural audio was recorded, with an average duration of 3-5 minutes per recording, considering this is the most likely length that someone would record in everyday life. In general, the recording person was allowed to talk while recording, but try to minimize the amount of his own talking. Also, the recording person was required to not move much (body or head movement), to allow possible future use of spatial information present in binaural recordings. The equipment used for recording this specific dataset consists of binaural Soundman OKM II Klassik/studio A3 electret in-ear microphones and Roland Edirol R09 wave recorder using 44.1 kHz sampling rate and 24 bit resolution.

B. Annotation

Annotation of the recorded materials was done at two levels: acoustic scene annotation at recording level and detailed sound events annotation in each recording for a subset of the data. The acoustic scene categories were decided in advance.

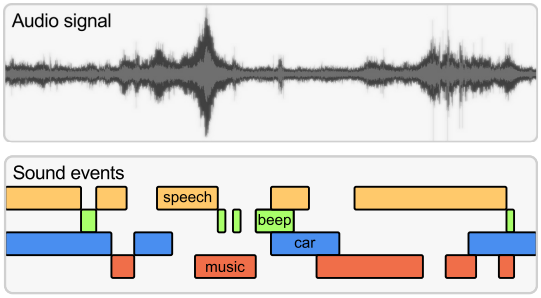

Fig. 1. Polyphonic annotation of audio.

Individual sound events in each recording were annotated using freely chosen labels for sounds. Nouns were used to characterize each sound source, and verbs to characterize the sound production mechanism, whenever this was possible. The ground truth is provided as a list of the sound events present in the recording, with annotated onset and offset for each sound instance. Sound events are overlapping, as illustrated in Fig. 1. Recording and annotation was done by two research assistants that were trained first on few example recordings. Each assistant annotated half of the data. They were instructed to annotate all audible sound events, and mark onset and offset as they consider fit. Because of the overlapping sounds, each recording had to be listened multiple times and therefore annotation was a very time consuming procedure.

C. Privacy screening and postprocessing

Postprocessing of the recorded and annotated data involves aspects related to privacy of recorded individuals, possible errors in the recording process, and analysis of annotated sound event classes. For audio material recorded in private places, written consent was obtained from all people involved. Material recorded in public places does not require such consent, but was screened for content, and privacy infringing segments were eliminated. Microphone failure and audio distortions were also annotated and this annotation is provided together with the rest of the data.

Analysis of sound event annotation reveals the diversity of the audio material. Labels for the sound classes were chosen freely, and this resulted in a large set of labels. There was no evaluation of inter-annotator agreement due to the high level of subjectivity inherent to the problem. Target sound event classes were selected based on the frequency of the obtained labels, to ensure that the selected sounds are common for an acoustic scene, and there are sufficient examples for learning acoustic models.

III. TUT Acoustic Scenes 2016

TUT Acoustic Scenes 2016 dataset consists of 15 different acoustic scenes: lakeside beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, urban park, residential area, train, and tram. All audio material was cut into segments of 30 seconds length.

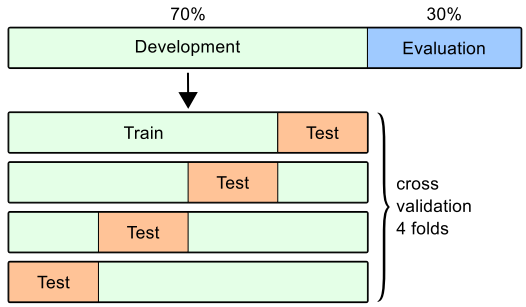

Fig. 2. Database partitioning into training and evaluation sets

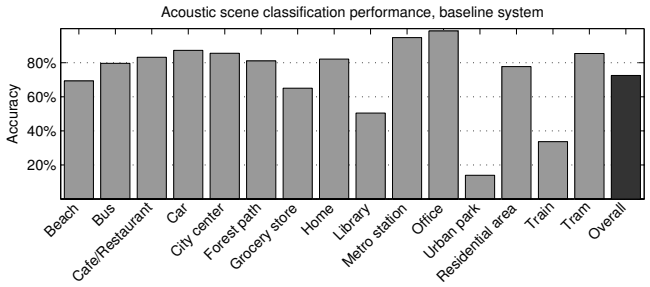

Fig. 3. TUT Acoustic Scenes 2016: Baseline system performance on development set

A. *Cross-validation setup*

The dataset was split into development set and evaluation set, such that the evaluation set consists 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. This process is illustrated in Fig. 2. 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 done based on the location of the original recordings. All segments obtained from the same original recording were included into a single subset - either development or evaluation. This is a very important detail that is sometimes neglected, and failing to recognize it results in overestimating the system performance, as the classification systems are capable of learning the location-specific acoustic conditions instead of the intended general audio scene properties,. The phenomenon is similar to the ”album effect” encountered in music information retrieval, that has been noticed and is usually accounted for when setting up experiments [19]. The cross-validation setup provided with the database consists of four folds distributing the 78 segments available in the development set based on location.

B. *Baseline system and evaluation*

The baseline system provided with the database consists of a classical mel frequency cepstral coefficient (MFCC) and Gaussian mixture model (GMM) based classifier. MFCCs were calculated for all audio using 40 ms frames with Hamming window and 50% overlap and 40 mel bands. The first 20 coefficients were kept, including the 0th order coefficient. Delta and acceleration coefficients were also calculated using a window length of 9 frames, resulting in a frame-based feature vector of dimension 60. For each acoustic scene, a GMM class model with 32 components was trained based on the described features using expectation maximization algorithm. The testing stage uses maximum likelihood decision among all acoustic scene class models. Classification performance is measured using accuracy: the number of correctly classified segments among the total number of test segments. The classification results using the cross-validation setup for the development set is presented in Fig. 3: overall performance is 72.5%, with context-wise performance varying from 13.9% for park to 98.6% for office.

IV. TUT Sound Events 2016

TUT Sound Events 2016 dataset consists of two common everyday environments: one outdoor - residential area - and one indoor - home. These are environments of interest in applications for safety and surveillance (outside home) and human activity monitoring or home surveillance. The audio material consists of the original full length recordings that are also part of TUT Acoustic Scenes.

Target sound event classes were selected based on the frequency with which they appear in the raw annotations and the number of different recordings they appear in. Mapping of the raw labels was performed, merging for example ”car engine running” to ”engine running”, and grouping various impact sounds with only verb description such as ”banging”, ”clacking” into ”object impact”.

The selected event classes and their frequency are listed in Table I. It can be observed that 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, besides some more well defined (still impact) dishes, cutlery, etc. The provided ground truth disregards all sound events that do not belong to the target classes, despite them being present in the audio. In this respect, we provide real-life audio with annotations for selected event classes. For completeness, the detailed annotations containing all available annotated sounds are provided with the data, but the sound event detection task is planned with the event classes presented in Table I.

TABLE I
TUT Sound Events 2016: Most frequent event classes and 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

A. Cross-validation setup

Partitioning of data into training and evaluation subsets was done based on the amount of examples available for each event class, while also taking into account recording location. Ideally the subsets should have the same amount of data for each class, or at least the same relative amount, such as a 70-30% split. Because the event instances belonging to different classes are distributed unevenly within the recordings, we can only control to a certain extent the partitioning of individual classes. For this reason, the condition was relaxed to including 60-80% of instances of each class into the development set for residential area, and 40-80% for home. The available recordings were repeatedly randomly assigned to the sets until this condition was met for all classes.

The development set was further partitioned into four folds, such that each recording is used exactly once as test data. At this stage the only condition imposed was that the test subset does not contain classes unavailable in training. Residential area sound events data consists of five recordings in the evaluation set and four folds distributing 12 recordings into training and testing subsets. Home sound events data consists of five recordings in the evaluation set and four folds distributing 10 recordings into training and testing subsets.

B. Baseline system and evaluation

The baseline system is based on MFCCs and GMMs, with MFCCs calculated using the same parameters as in the acoustic scenes baseline system. For each event class, a binary classifier was set up. The class model was trained using the audio segments annotated as belonging to the modeled event class, and a negative model was trained using the rest of the audio. In the test stage, the decision is based on likelihood ratio between the positive and negative models for each individual class, with a sliding window of one second.

Evaluation of system performance for sound event detection uses error rate and F-score in a fixed time grid, as defined in [20]. In segments of one second length, the activities of sound event classes are compared between the ground truth and the system output. An event is considered correctly detected in a given segment if both the ground truth and system output indicate it as active in that segment. Other case are: false positive if the ground truth indicates an event as inactive and the system output indicates it as active, false negative if the ground truth indicates it as active and the system output indicates it as inactive.

Based on the total counts of true positives TP, false positives FP and false negatives FN, precision, recall, and F-score are calculated according to the formula:

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \quad F = \frac{2PR}{P+R} \qquad (1)$$

Error rate measures the amount of errors in terms of *insertions* (I), *deletions* (D) and *substitutions* (S). A substitution is defined as the case when the system detects an event in a given segment, but gives it a wrong label. This is equivalent to the system output containing one false positive and one false negative in the same segment. After counting the number of substitutions per segment, the remaining false positives in the system output are counted as insertions, and the remaining false negatives as deletions. The error rate is then calculated by integrating segment-wise counts over the total number of segments K, with $N(k)$ being the number of active ground truth events in segment k [21]:

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)} \qquad (2)$$

Event-based metrics consider true positives, false positives and false negatives with respect to event instances. An event in the system output is considered correctly detected if it has a temporal position overlapping with the temporal position of an event with the same label in the ground truth. A collar of 200 ms was allowed for the onset, and for offset either the same 200 ms collar or a tolerance of 50% with respect to the ground truth event duration. An event in the system output that has no correspondence to an event with same label in the ground truth within the allowed tolerance is a false positive, and an event in the ground truth that has no correspondence to an event with same label in the system output within the allowed tolerance is a false negative. Event-based substitutions are defined differently than segment-based: events with correct temporal position but incorrect class label are counted as substitutions, while insertions and deletions are the events unaccounted for as correct or substituted in system output or ground truth, respectively. Precision, recall, F-score and error rate are defined the same way, with error rate being calculated with respect to the total number of events in the ground truth.

TABLE II
TUT SOUND EVENTS 2016: BASELINE SYSTEM PERFORMANCE ON DEVELOPMENT SET

Acoustic scene	Segment-based		Event-based	
	ER	F [%]	ER	F [%]
home	0.95	18.1	1.33	2.5
residential area	0.83	35.2	1.99	1.6
average	0.89	26.6	1.66	2.0

Performance of the baseline system on the training and development subset is presented in Table II. The results from all folds were combined to produce a single evaluation, for avoiding biases caused by data imbalance between folds [22]. While the segment-based performance is not discouraging, the performance of this baseline system evaluated using event-based metrics is extremely poor. This is easily explained by the fact that the system does not use any specific segmentation step, and it relies on the classifier to decide activity of sound classes. The binary classification scheme is not capable of detecting onsets and offsets within the evaluated tolerance. An error rate over 1.0 is also an indication of the system producing more errors than correct outputs.

A closer inspection of segment-based results reveals that there is big difference in the capability of the system to detect different classes. As can be seen in Table III, some classes are

TABLE III
TUT SOUND EVENTS 2016: SEGMENT-BASED F-SCORE CALCULATED CLASS-WISE

Residential area		**Home**	
event class	F [%]	event class	F [%]
(object) banging	0.0	(object) rustling	8.3
bird singing	33.8	(object) snapping	0.0
car passing by	59.9	cupboard	0.0
children shouting	0.0	cutlery	0.0
people speaking	30.6	dishes	4.3
people walking	2.8	drawer	8.1
wind blowing	14.2	glass jingling	0.0
		object impact	22.8
		people walking	18.3
		washing dishes	24.6
		water tap running	41.2

correctly detected in about a third of the segments, while for car passing by the detection rate is over 50%. On the other hand, the system completely fails to detect some classes. This is not surprising, considering the simplicity of the system.

V. CONCLUSIONS AND FUTURE WORK

In this paper we introduced a dataset for acoustic scene classification and sound event detection in real-world audio. The development set for both is currently available for download [23], [24], while the evaluation set will be published soon. The provided database is more complex in terms of sound event classes than previous ones, and was carefully collected to obtain a high acoustic variability of acoustic scenes. We recommend the use of the cross-validation setup for publishing future results, as this will allow exact comparison between systems. The provided cross-validation setup also ensures that all audio recorded at the same location is placed in the same subset, such that there is no data contamination between training and testing sets.

Future work will extend this data in both acoustic scenes and sound events. Other teams are invited to contribute to the dataset, by using same recording and annotation principles. The annotation procedure will be developed to improve annotation speed and and avoid ambiguity in sound event labels. Additionally, inter-annotator agreement can be used to combine the output from multiple annotators to minimize as much as possible the subjectivity of the ground truth.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND.

REFERENCES

[1] D. Battaglino, L. Lepauloux, L. Pilati, and N. Evans, "Acoustic context recognition using local binary pattern codebooks," in *Worshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, October 2015.

[2] 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 Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, jan 2006.

[3] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[4] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene detection," Tech. Rep., HAL, 2014.

[5] V. Bisot, S. Essid, and G. Richard, "Hog and subband power distribution image features for acoustic scene classification," in *2015 European Signal Processing Conference (EUSIPCO)*, Nice, France, August 2015, pp. 724–728.

[6] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *IEEE International Conference on Multimedia and Expo*, Los Alamitos, CA, USA, 2005, pp. 1306–1309, IEEE Computer Society.

[7] A. Härmä, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Los Alamitos, CA, USA, 2005, pp. 634–637, IEEE Computer Society.

[8] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, pp. 22 – 28, 2015.

[9] Ya-Ti Peng, Ching-Yung Lin, Ming-Ting Sun, and Kun-Cheng Tsai, "Healthcare audio event classification using hidden markov models and hierarchical hidden markov models," in *IEEE International Conference on Multimedia and Expo*, June 2009, pp. 1218–1221.

[10] S. Goetze, J. Schröder, S. Gerlach, D. Hollosi, J.E. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering*, vol. 6, no. 1, pp. 40–50, March 2012.

[11] P. Guyot, J. Pinquier, X. Valero, and F. Alias, "Two-step detection of water sound events for the diagnostic and monitoring of dementia," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, July 2013, pp. 1–6.

[12] D. Stowell and D. Clayton, "Acoustic event detection for multiple overlapping similar sources," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustic (WASPAA)*, New Paltz, NY, October 2015.

[13] R. Cai, Lie Lu, A. Hanjalic, Hong-Jiang Zhang, and Lian-Hong Cai, "A flexible framework for key audio effects detection and auditory context inference," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 1026–1039, May 2006.

[14] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 4, no. 2, pp. 1–23, 2008.

[15] M. Bugalho, J. Portelo, I. Trancoso, T.s Pellegrini, and A. Abad, "Detecting audio events for semantic video search.," in *Interspeech*, 2009, pp. 1151–1154.

[16] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech and Music Processing*, 2013.

[17] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M.D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013*, Oct 2013, pp. 1–4.

[18] "Detection and Classification of Acoustic Scenes and Events 2016, IEEE AASP Challenge," http://www.cs.tut.fi/sgn/arg/dcase2016/, [Online; accessed 5-Feb-2016].

[19] Y.E. Kim, D.S. Williamson, and S. Pilli, "Towards quantifying the album-effect in artist classification," in *In Proceedings of the International Symposium on Music Information Retrieval*, 2006.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162, 2016.

[21] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 048317, 2007.

[22] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement," *SIGKDD Explor. Newsl.*, vol. 12, no. 1, pp. 49–57, Nov. 2010.

[23] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Acoustic Scenes 2016," https://zenodo.org/record/45739, 2016, DOI: 10.5281/zenodo.45739.

[24] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Sound Events 2016," https://zenodo.org/record/45759, 2016, DOI: 10.5281/zenodo.45759.