MAVD: A DATASET FOR SOUND EVENT DETECTION IN URBAN ENVIRONMENTS

Pablo Zinemanas, Pablo Cancela, Martín Rocamora

Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay
{pzinemanas, pcancela, rocamora}@fing.edu.uy

ABSTRACT

We describe the public release of a dataset for sound event detection in urban environments, namely MAVD, which is the first of a series of datasets planned within an ongoing research project for urban noise monitoring in Montevideo city, Uruguay. This release focuses on traffic noise, MAVD-traffic, as it is usually the predominant noise source in urban environments. An ontology for traffic sounds is proposed, which is the combination of a set of two taxonomies: vehicle types (e.g. car, bus) and vehicle components (e.g. engine, brakes), and a set of actions related to them (e.g. idling, accelerating). Thus, the proposed ontology allows for a flexible and detailed description of traffic sounds. We also provide a baseline of the performance of state–of–the–art sound event detection systems applied to the dataset.

Index Terms— SED database, traffic noise, urban sound

1. INTRODUCTION

Recent years have witnessed the upsurge of the Smart City concept, i.e. networks of Internet of Things (IoT) sensors used to collect data in order to monitor and manage city services and resources. Noise levels in cities are often annoying or even harmful to health, being consequently among the most frequent complaints of urban residents [1]. This fuelled the development of technologies for monitoring urban sound environments, mainly oriented towards the mitigation of noise pollution [2, 3]. The application of signal processing and machine learning has lead to the automatic generation of high–level descriptors of the sound environment. This encompasses the problem of sound event detection (SED), as an attempt at describing the acoustic environment through the sounds encountered in it. It is defined as the task of finding individual sound events, by indicating the onset time, the duration and a text label describing the type of sound [4, 5].

The SED problem is usually approached within a supervised learning framework, using a set of predefined sound event classes and annotated audio examples of them [5, 6]. One of the most challenging aspects of the problem is that it involves the detection of overlapping sound events. In addition, given the intrinsic variability of sound sources of the same type (e.g. cars) and the influence of the acoustic environment (e.g. reverberation, distance) for different locations and situations, the acoustic features of each class can exhibit great diversity. The solutions proposed typically use a mel–spectral representation of the audio signal as the input features, and apply different classification methods, including Random Forest [7], GMM [8], and more recently convolutional neural networks [9, 10] and recurrent neural networks [11, 12, 13].

1.1. Related work

Publicly available datasets for SED are of crucial importance to foster the development of the field as they encourage reproducible research and fair comparison of algorithms. In this respect, the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge, held for the first time in 2013 and repeated every year since 2016, has established a benchmark for sound event detection using open data [6, 14].

Two of the datasets used in the DCASE challenge for SED in urban environments are part of the TUT database (TUT Sound Events 2016 and 2017), which was collected in residential areas in Finland by Tampere University of Technology (TUT) and contain overlapping sound events manually annotated [8]. The classes are defined during the labeling process. In a first step, the participants are asked to mark all the sound events freely, and later the labels are grouped into more general concepts. In addition, the tags must be composed of a noun and a verb, such as ENGINE ACCELERATING [8].

Manual annotation of audio recordings for SED is a very time consuming task, primarily due to multiple overlapping sounds, which has limited the amount of annotated audio available. A way to alleviate the work involved in manual annotation is to use weak labels, as in DCASE 2017 task 4 [14], which indicate the presence of a source without giving time boundaries. Another approach is to create synthetic audio mixtures using isolated sound events. This is the approach adopted in the URBAN-SED dataset [9], that contains synthesized soundscapes with sound event annotations generated using Scaper [9] (a software library for soundscape synthesis). The original sound events are extracted from the UrbanSound8K dataset [15], where a taxonomic categorization of urban sounds is proposed. At the top level, four groups are defined: HUMAN, NATURAL, MECHANICAL and MUSICAL, which have been used in previous works. To define the lower levels, the most frequent noise complaints in New York city from 2010 to 2014 were used [15].

Table 1 summarizes the characteristics of the available datasets for SED in urban environments. While the TUT datasets are limited to only one and two hours, the URBAN-SED dataset comprises 30 hours of audio but contains synthetic audio mixtures instead of real recordings. Other resources for research on urban sound environments are available, such as the SONYC Urban Sound Tagging (SONYC–UST) dataset [2], though they are not specifically devoted to the SED problem. If traffic sounds are to be considered, the DCASE 2017 task 4 training dataset has only weak labels, the TUT database has only a moderate amount of traffic activity since it was recorded in a calm residential area, and only three out of the ten classes in URBAN-SED are related to traffic (i.e. CAR_HORN, ENGINE_IDLING and SIREN). Therefore, there is plenty of room for expanding the existing resources, in particular, for specific applications' scenarios such traffic noise monitoring.

https://doi.org/10.33682/kfmf-zv94

dataset	classes	hours	type	label
TUT-SE 2016 [8]	7	1	recording	strong
TUT-SE 2017 [8]	6	2	recording	strong
URBAN-SED [9]	10	30	synthetic	strong
DCASE2017 #4 [14]	17	141	10-s clips	weak
MAVD-traffic	21	4	recording	strong

Table 1: Available datasets for SED in urban environments, along with the released dataset.

1.2. Our contributions

We describe the first public release of a dataset for SED in urban environments, called MAVD, for Montevideo Audio and Video Dataset. This release focuses on traffic sounds, namely MAVD-traffic, which corresponds to the most prevalent noise source in urban environments. The records were generated in various locations in Montevideo city and include both audio and video files, along with annotations of the sound events. The video files, apart from being useful for manual annotation, open up new research possibilities for SED using audio and video. The annotations follow a new ontology for traffic sounds that is proposed in this work. It arises from the combination of a set of two taxonomies: vehicle types (e.g. car, bus) and vehicle components (e.g. engine, brakes), and a set of actions related to them (e.g. idling, accelerating). Thus, the proposed ontology allows for a flexible and detailed description of traffic sounds. In addition, we provide a baseline of the performance of state–of–the–art SED systems applied to the MAVD-traffic dataset. Finally, we discuss possible directions for further research and some efforts we undertaken to improve and extend current dataset.

2. ONTOLOGY

The proposed ontology focuses on traffic noise. Consequently, vehicles (such as cars, buses, motorcycles and trucks) are the main sources of noise and define the classes of interest. However, vehicles generate different types of sounds, for example those related to the braking system, the rolling of the wheels or the engine, calling for a classification that is more specific than just the type of vehicle. One way to approach it, is by classifying sound events with different correlated attributes, such as the sound source (object), the action, and the context [16]. These attributes can be defined by one or several taxonomies, implying that the same event can be classified by several schemes simultaneously [16]. In this case, the context is defined by urban environments where traffic noise is predominant. Then, sound sources and actions can be described by several taxonomies, for instance, one that defines the type of vehicle and other that defines the internal components that generate the sound.

We define an ontology based on a graph like the one shown in Figure 1, which consists of two taxonomies that blend in the middle: the top one describes the categories of vehicles; and the bottom one describes the categories of components. The categories of components are further combined with a set of actions to form an object-action pair (e.g. ENGINE IDLING, ENGINE ACCELERATING).[1]

The categories indicated in bold are those that are called *basic level* (CAR, BUS, etc. for vehicles and ENGINE, WHEEL, etc. for components). These two taxonomies of the ontology are merged into what is called the *subordinate level* (depicted in italics), which are combinations of elements of the categories of vehicles and components, with the aim of providing a more detailed description of the noise source (e.g. CAR/ENGINE IDLING, BUS/COMPRESSOR). Note that the diagram of Figure 1 does not show all the class labels.

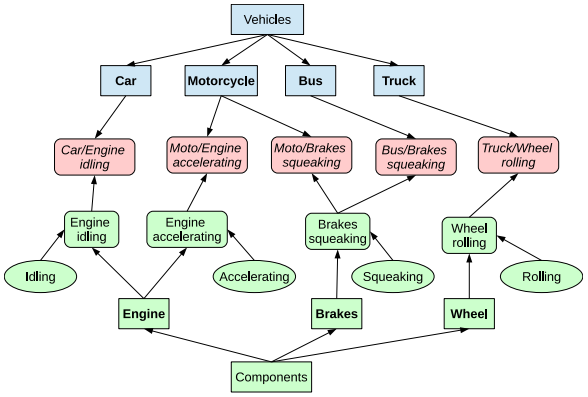

Figure 1: Graph representing the ontology. The top taxonomy refers to the vehicle categories and the bottom one to the components. The basic levels are indicated in bold and the subordinate level is marked in italics. The rectangle nodes denote objects; the ellipses denote actions; and the rounded rectangles indicate objects–actions pairs.

3. DATASET

3.1. Recordings

The recordings were produced in Montevideo, the capital city of Uruguay, which has population of 1.4 million people. Four different locations were included in this release of the dataset, corresponding to different levels of traffic activity and social use characteristics:

L1. Residential area, with several shops and many buses.

L2. Park area. No housing or shops. Some light traffic nearby.

L3. Park/residential area. Similar to location L2, but next to a residential area, with more traffic noise and less nature sounds.

L4. Residential area, with a few shops and some buses.

The sound was captured with a SONY PCM-D50 recorder at a sampling rate of 48 kHz and a resolution of 24 bits. The video was recorded with a GoPro Hero 3 camera at a rate of 30 frames per second and a resolution of 1920×1080 pixels. Audio and video files of about 15–minutes long were recorded at different times of the day in the different locations.

Some basic processing was done to generate the files of the dataset from the raw recordings. This included the synchronization of audio and video, the removal of windy sections and the segmentation into excerpts of approximately five minutes to facilitate their manipulation. The train and validation sets are composed of 24 and 7 files from the location L1 respectively, while the test set consists of 16 files from the L2, L3 and L4 locations[2]. The dataset totals 233 minutes (almost 4 hours, as shown in Table 1), of which 117 minutes correspond to the train set, 33 minutes to the validation set and 83 minutes to the test set.

[1] This could also be done in the top taxonomy for the vehicles, for example BUS PASSING BY, CAR STOPPING, etc., but was considered redundant.

[2] In train/validation we favoured the location with more events (L1) but other fold schemes could be implemented using the metadata information.

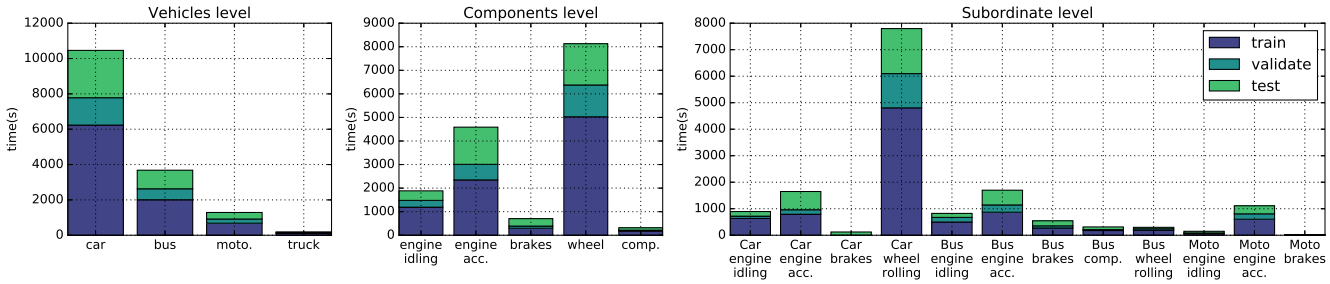

Figure 2: Total time for each class in the dataset. The first two graphs correspond to the basic levels, and the third one to the subordinate level.

3.2. Annotation

The ELAN [17] software was used to manually annotate the recordings of the dataset. The software allows the user to simultaneously inspect several audio and/or video recordings and produce annotations time–aligned to the media. During the annotation process the software session displayed the audio waveform, the video record and the spectrogram of the audio signal. For the latter, an auxiliary video file was generated for each recording, showing the spectrogram of the audio signal and a vertical line indicating current time instant (as shown in Figure 3). The annotations can be created on multiple layers, which can be hierarchically interconnected. This feature is a perfect fit for the taxonomies' approach defined above.

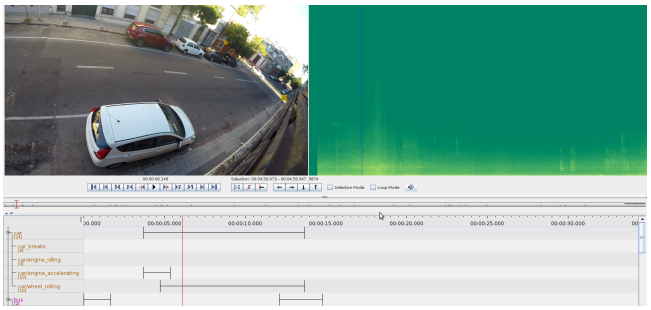

Figure 3: Screenshot of the ELAN software showing MAVD–traffic data: at the top the video and the spectrogram (with a marker indicating the instant being labeled) and at the bottom the annotations.

The annotation process was carried out in two steps. First, the vehicle categories were labeled (e.g. CAR, BUS). Then, for each of the marked segments, the labels of the component categories (e.g. ENGINE IDLING) were annotated to form the subordinate level. Figure 2 shows the total duration of the events for the three category types. Note that the dataset is highly unbalanced, especially the subordinate level, being CAR/WHEEL ROLLING the predominant class.

4. EXPERIMENTS AND RESULTS

4.1. Experiments

We devised two different experiments to provide a baseline of the SED performance on the MAVD–traffic dataset. For the first experiment, we used a Random Forest classifier with the acoustic features defined as follows. We extracted 20 mel–frequency cepstral coefficients (MFCC) using the energy in 40 mel bands. The MFCCs were calculated in frames of 40 ms overlapped 50% and using a Hamming analysis window. Besides, first and second derivatives were calculated (ΔMFCC, Δ^2MFCC), to describe the temporal variations of the coefficients. The features were computed with *librosa* (version 0.6.1) [18] and the Random Forest models were implemented with *scikit-learn* (version 0.17) [19].

For the second experiment, we used the convolutional neural network for SED proposed by Salamon et. al in [9] (S–CNN). The input of this network is a one–second length mel–spectrogram and has three convolutional layers followed by three fully–connected layers. The final layer is a sigmoid that performs the classification task (the number of units is equal to the number of classes). First we trained the S–CNN model with the URBAN–SED dataset using the same strategy used in [9]. Then, we used a fine–tuning strategy in order to specialize the network to the MAVD–traffic dataset. We replace the last sigmoid layer of the network to accomplish the classification task of the MAVD-traffic dataset. The parameters of the other layers of the network were kept unchanged during the fine–tuning training process. The S–CNN model was implemented in *keras* (version 2.2.0) [20] using *tensorflow* (version 1.5.0) [21].

4.2. Metrics

The performance measures typically used for the SED problem are: F–score ($F1$) and Error Rate (ER), on a fixed time grid [22]. The detected sound events are compared with the *ground–truth* in one–second length segments. Based on the number of false positives (FP) and false negatives (FN), the values of the *precision* (P) and *recall* (R) are computed. Then, the F–score ($F1$) is calculated as:

$$F1 = \frac{2PR}{P+R} = \frac{2TP}{2TP+FN+FP}. \quad (1)$$

The error rate (ER) is calculated in terms of insertions $I(k)$, deletions $D(k)$ and substitutions $S(k)$ in each segment k. A substitution is defined as the case in which the system detects an event in a segment but with the wrong label. This corresponds to a simultaneous FP and FN for the segment. The remaining FP not included in the substitutions are considered insertions and the remaining FN as deletions. Finally, the ER is calculated considering all errors as:

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

where K is the total number of segments and $N(k)$ is the number of active classes in the *ground-truth* at segment k [8, 22].

The values of $F1$ and ER are usually calculated globally over the full set of segments and classes simultaneously. They can also

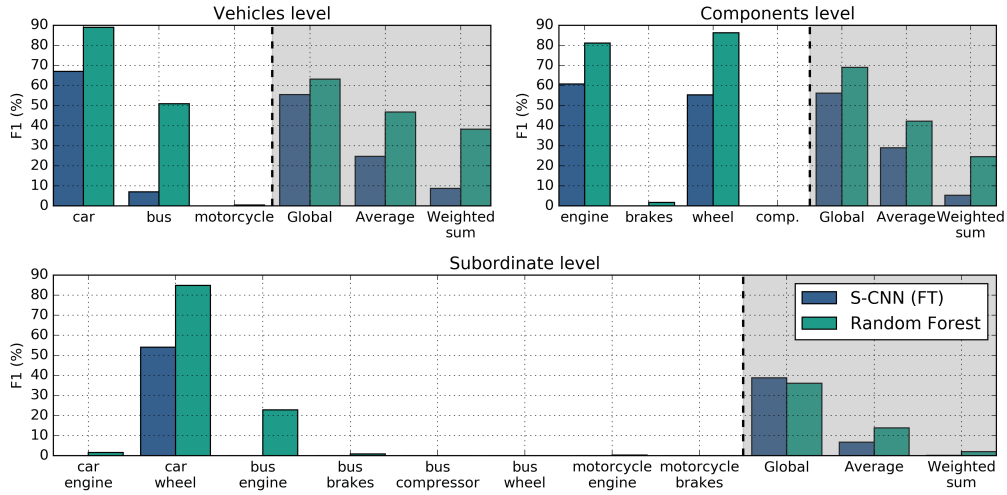

Figure 4: Comparison of the SED results for both S-CNN and Random Forest systems applied to the MAVD–traffic dataset. The performance is shown at the basic and subordinate levels for the different classes and for the three discussed metrics: Global, Average and Weighted sum.

be calculated restricted to each class and then averaged, which are denoted as $\bar{F}1$ and $\bar{ER}$ respectively. This average is calculated as:

$$\bar{M} = \frac{1}{C}\sum_{c=1}^{C} M_c, \tag{3}$$

where C is the number of classes and M_c is the metric for class c.

These global metrics can bias the SED algorithms to detect only the majority class. This is illustrated by the results of DCASE challenges 2016 and 2017, in which the algorithms that obtained better global results actually detect only the majority class [6, 14].

We aim to improve these evaluation metrics (ER and $F1$) in the case of multi–class SED systems trained with very unbalanced data, by increasing the importance of detecting the minority classes. To do so, we propose a weighted sum of the metric values as,

$$\hat{M} = \sum_{c=1}^{C} w_c M_c, \qquad w_c = \frac{1/N_c}{\sum_{j=1}^{C} 1/N_j} \tag{4}$$

where N_c is the number of active segments for class c, and w_c is the weight for each class, which is designed to give more importance to the minority classes. Note that w_c increases when N_c decreases, as expected. The sum in the denominator ensures that $\sum_c w_c = 1$.

4.3. Results

We trained the Random Forest and the S–CNN models for the three class levels (vehicles, components and subordinate) and obtained the results shown in Figure 4 and in Table 2. Note that the S–CNN models tend to classify only the majority class while yielding quite good results for the global ER and $F1$ metrics, as discussed in Section 4.2. On the other hand, the weighted sum metrics, $\hat{ER}$ and $\hat{F}1$, clearly penalize the detection of only the majority class. The Random Forest models perform better in detecting the minority classes (see the BUS class), reaching higher values of the weighted sum metrics. The source code for training the models and reproducing these results on the MAV-traffic dataset is publicly available.[3]

[3] https://github.com/pzinemanas/MAVD-traffic

Level	Model	Global ER	Global $F1(\%)$	Weighted sum $\hat{ER}$	Weighted sum $\hat{F}1(\%)$
Vehicles	RF	0.54	63.1	0.71	38.2
	S–CNN	0.51	55.5	0.97	8.70
Components	RF	0.49	69.0	0.80	24.6
	S–CNN	1.17	56.2	1.03	5.35
Subordinate	RF	0.78	36.1	0.96	1.98
	S–CNN	0.70	38.9	1.00	0.17

Table 2: Results for Random Forest (RF) and S-CNN using the original (Global) and the proposed (Weighted sum) metrics.

5. CONCLUSION

In this work a new dataset for SED in urban environments is described and publicly released.[4] The dataset focuses on traffic noise and was generated from real recordings in Montevideo city. Apart from audio recordings it, also includes synchronized video files.[5] The dataset was manually annotated using an ontology proposed in this work, which combines two taxonomies (vehicles and component–action pairs) for a detailed description of traffic noise sounds. Since the taxonomies follow a hierarchy they can be used with different levels of detail. The performance of two SED system is reported as a baseline for the dataset. Some considerations are given regarding the evaluation metrics for class–unbalanced datasets. In future work, we will increase the size of the dataset, by including other locations with different levels of traffic activity. We also plan to address urban soundscapes in which other noise sources are predominant, such as those related to social, construction or industrial activities. In addition, image processing techniques will be applied to the video files to develop a multi–modal SED system.

[4] Available from Zenodo, DOI 10.5281/zenodo.3338727

[5] In this release, the video files are available in low resolution as we are anonymizing them, after which they will be available in high resolution.

6. REFERENCES

[1] H. Ising and B. Kruppa, "Health effects caused by noise: Evidence in the literature from the past 25 years," *Noise & health*, vol. 6, pp. 5–13, 11 2004.

[2] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, Feb 2019.

[3] D. K. Daniel Steele and C. Guastavino, "The sensor city initiative: cognitive sensors for soundscape transformations," in *Geoinformatics for City Transformations*. Technical University of Ostrava, January 2013, pp. 243–253.

[4] J. P. Bello, C. Mydlarz, and J. Salamon, *Computational Analyis of Sound Scenes and Events*. Springer, 2017, ch. 13 Sound Analysis in Smart Cities.

[5] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analyis of Sound Scenes and Events*. Springer, 2017, ch. 1 Introduction to Sound Scene and Event Analysis.

[6] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[7] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experimentation on the dcase challenge 2016: Task 1 - acoustic scene classification and task 3 - sound event detection in real life audio," in *Detection and Classification of Acoustic Scenes and Events 2016*, 2016.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *In 24rd European Signal Processing Conference 2016 (EUSIPCO 2016), Budapest, Hungary*, 2016.

[9] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello., "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, october 2017.

[10] I.-Y. Jeong, S. Lee, Y. Han, and K. Lee, "Audio event detection using multiple-input convolutional neural network," DCASE2017 Challenge, Tech. Rep., September 2017.

[11] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Detection and Classification of Acoustic Scenes and Events 2016*, 2016.

[12] R. Lu and Z. Duan, "Bidirectional GRU for sound event detection," DCASE2017 Challenge, Tech. Rep., September 2017.

[13] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *Transactions on Audio, Speech and Language Processing: Special issue on Sound Scene and Event Analysis*, vol. 25, no. 6, pp. 1291–1303, June 2017.

[14] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, June 2019.

[15] 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)*, Orlando, FL, USA, Nov. 2014.

[16] C. Guastavino, *Computational Analyis of Sound Scenes and Events*. Springer, 2017, ch. 7 Everyday Sound Categorization.

[17] Max Planck Institute for Psycholinguistics, "ELAN - the lenguage active." [Online]. Available: tla.mpi.nl/tools/tla-tools/elan/

[18] B. McFee, M. McVicar, C. Raffel, D. Liang, O. Nieto, E. Battenberg, J. Moore, D. Ellis, R. Yamamoto, R. Bittner, D. Repetto, P. Viktorin, J. F. Santos, and A. Holovaty, "librosa: 0.4.1," Oct. 2015. [Online]. Available: https://doi.org/10.5281/zenodo.32193

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[21] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

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