CROWDSOURCING STRONG LABELS FOR SOUND EVENT DETECTION

Irene Martín-Morató, Manu Harju, Annamaria Mesaros

Computing Sciences, Tampere University
Korkeakoulunkatu 7, 33720 Tampere, FINLAND
{irene.martinmorato, manu.harju, annamaria.mesaros}@tuni.fi

ABSTRACT

Strong labels are a necessity for evaluation of sound event detection methods, but often scarcely available due to the high resources required by the annotation task. We present a method for estimating strong labels using crowdsourced weak labels, through a process that divides the annotation task into simple unit tasks. Based on estimations of annotators' competence, aggregation and processing of the weak labels results in a set of objective strong labels. The experiment uses synthetic audio in order to verify the quality of the resulting annotations through comparison with ground truth. The proposed method produces labels with high precision, though not all event instances are recalled. Detection metrics comparing the produced annotations with the ground truth show 80% F-score in 1 s segments, and up to 89.5% intersection-based F1-score calculated according to the polyphonic sound detection score metrics.

***Index Terms*—** Strong labels, Sound event detection, Crowdsourcing, Multi-annotator data

1. INTRODUCTION

The research on sound event detection is currently dominated by methods that learn acoustic models from weakly-labeled data [1, 2], in which only presence of sound events is indicated, without explicit temporal information. However, the task of sound event detection is defined as recognizing and temporally locating sound instances within a recording [3], which creates a mismatch between the training and the requirements imposed on the system. The main cause of this situation is the lack of suitable datasets containing strong labels, that indicate both textual labels and temporal boundaries of events.

While training of acoustic models can be achieved with advanced learning methods using weakly-labeled real-life recordings or strongly-labeled synthetic audio mixtures, the evaluation of such methods still requires strongly-labeled data. Typically, a small amount of data is manually annotated for evaluation. In recent data evaluation challenges, the evaluation data consisted of short recordings, of length 10 seconds [2], while earlier ones used recordings of 3-5 minutes [4]. Such annotation efforts were concentrated to individual research groups, resulting in datasets of small size. The most recent work introduces strong labels for a part of AudioSet [5], providing 67k manually annotated clips. The length of these clips is 10 s, and they contain on average 3.5 labels [6].

Manual annotation of sound events is subjective in many ways, from the textual labels selected for the sound [7], to the placing of temporal boundaries for the event instances [3]. Ideally, an objective reference annotation is based on multiple annotators or curation, but curation of strong labels is difficult. For example in [6], a first-pass labeling was reviewed by a different annotator who could modify the labels, but even with 5 stages this process rarely converged to consensus. On the other hand, multiple, independent annotators should be followed by a method for aggregating the information. In image analysis, aggregation is usually done as intersection of segments [8] or by maximizing agreement between annotators [9]. For audio, it is rare to have data with multiple annotators; in particular, strong labeling is impractical for multi-annotator solutions because of the difficulty of the task.

An attractive solution for annotating large amounts of data, increasingly used for audio annotation, is crowdsourcing [10, 11, 12]. Crowdsourcing is suitable for simple annotation tasks like classification, with a number of services offering ready-made solutions for it, but these are not suitable in their current form for strong labeling.

In this paper, we propose a method for creating strong labels based on weak labels of overlapping segments. Weak labeling allows use of crowdsourcing, facilitating annotation of high volume of audio files in a fast way. The novelty of this work is threefold: 1. We introduce a novel, proof of concept, method for crowdsourcing strong labels, by estimating the strong labels based on audio tags (weak labels); 2. We evaluate the correctness of the collected annotation with respect to the ground truth, by using synthetically generated audio mixtures for which the reference annotation is generated at the same time with the audio mixtures; 3. We investigate the effect of reference annotations on the evaluated performance of sound event detection systems by using the ground truth annotations in training, but evaluating against the manually created reference annotation.

We show that it is possible to obtain reasonable strong labels by crowdsourcing segment-level tags and further processing them. The resolution of the estimated labels is determined by the degree of overlap of consecutive segments. We use annotator competence estimation [13] to eliminate the poor quality answers from the collected data before further processing. We also show that the mismatch between the synthetic ground truth and the manually created annotation produces a significant drop in measured performance, even though it does not affect system functionality. All data produced and collected in this study is publicly available.[1] The paper is organized as follows: Section 2 presents the proposed annotation procedure and data processing to estimate the strong labels. Section 3 presents the experimental setup, dataset, annotattion process, and analysis of the collected annotations. Section 4 shows the use of crowdsourced annotations in evaluation of a sound event detection system. Finally, Section 5 presents conclusions and future work.

This paper has received funding from Academy of Finland grant 332063 "Teaching machines to listen".

[1] MAESTRO Synthetic, 10.5281/zenodo.5126478

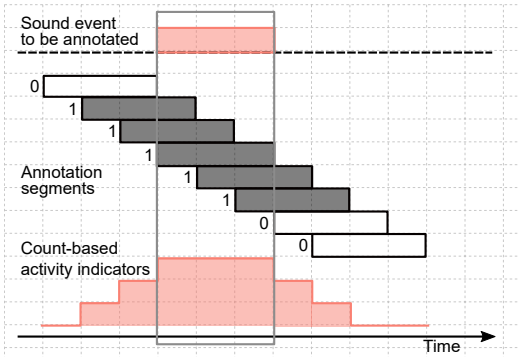

Figure 1: Estimating event activity from overlapping weakly-labeled segments.

2. ANNOTATION METHOD AND PROCESSING

2.1. Annotation method

In order to elicit a consistent behavior from annotators and produce a consistent output, an annotation task should require a single, simple decision [3]. Annotating audio with strong labels by requiring from the annotator textual labels for sounds, along with onset and offset for each sound instance, is the exact opposite. We propose a procedure that divides the strong annotation task into unit tasks that involve simple decision-making: to indicate presence of sounds from a pre-selected list labels in pre-segmented audio [3], practically weakly labeling individual segments. In addition to simplifying the work of the annotator, this approach makes the annotation task suitable for crowdsourcing. The audio files are segmented into short, overlapping segments, which are then annotated with weak labels by indicating binary activity of sound events within the entire segment. The list of target sound classes is selected in advance and presented to the annotator. The proposed method, illustrated in Fig. 1, uses a sliding "annotation window" over the length of the audio file, with a high rate of overlap between consecutive segments. The temporal activity of sounds within the original long file can then be estimated based on the tags of the individual segments, by reconstructing the temporal sequence of these segments into an aggregated representation that counts activity indicators at each time step. If all annotations are correct, event boundaries correspond to the boundaries of the maximum-valued region in the count-based activity indicators.

In this work, we choose a segment length of 10 seconds. We rely for this choice on studies that show accurate recognition of sound sources to be at most 6.8 seconds for a list of 42 different sounds [14]. With a hop of one second between the segments, the temporal reconstruction of the events activity will have a one second resolution, similar to the diffuse labels created in [6]. Following the procedure from [10], which was also used in [15], the task is presented as a single-pass multi-label annotation, in which the presence of a sound is explicitly indicated, and the absence is implicit by the label not being selected.

2.2. Annotator competence and ground truth estimation

Multiple annotator opinions are typically aggregated using majority vote in order to estimate reference labels for the data [10, 12]. In this work we use MACE - Multi-Annotator Competence Estimation [13] to identify the trustworthy annotators and to predict the labels. MACE models the behavior of the annotators in order to estimate the competence of the annotators and the true labels of the data. Annotator behavior is modeled using a binary variable drawn from a Bernoulli distribution, encompassing the annotator trustworthiness and the spamming behavior. The true labels and the spamming indicators are estimated based on the observed annotations, using expectation maximization. For complete details on the model assumptions and the method, we refer the reader to [13]. In the estimation of the true labels, the annotators' opinions are weighted based on their competence, in contrast to majority voting where each annotator's opinion has the same weight. The method has been extended for the audio tagging scenario in our previous work, by considering each multi-labeled item as a set of binary *yes/no* labels per item [15]. Each (item, label) pair is used as a separate annotator opinion, for which MACE estimates the true label. This approach models the single-pass multi-label annotation as a multiple-pass binary annotation [10]. The method was shown to recognize well the spamming behavior of annotators in audio tagging [15], providing a satisfactory level of inter-annotator agreement when the least competent annotators' opinions are removed from the data pool.

event class	instances
car horn	109
children voices	236
dog bark	343
engine idling	564
siren	256
street music	89

Table 1: Number of instances of each event class in the data

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Dataset and annotation procedure

Audio files are generated using Scaper [16], with small changes to the synthesis procedure. A soundscape is generated by placing events iteratively at random intervals until the desired maximum polyphony of 2 is obtained. Intervals between two consecutive events are selected at random, but limited to 2-10 seconds. Event classes and event instances are chosen uniformly, and mixed with a signal-to-noise ratio (SNR) randomly selected between 0 and 20 dB over a Brownian noise background. Having two overlapping events from the same class is avoided. Foreground events are extracted from the UrbanSound dataset [17]. The dataset includes classes: car_horn, children_playing, dog_bark, engine_idling, siren, and street_music, with children_playing renamed to children_voices for the annotation task, and files shorter than one second or longer than 60 seconds discarded. The classes were selected to mimic the street scenes annotated in [15]. Dataset statistics are presented in Table 1.

The dataset consists of 20 generated soundscapes, each having a length of 3 minutes. The resulting files are cut into 10 second segments with 1 second offsets, resulting in 171 segments from a single 3 minute soundscape, and a total number of 3420 10-second clips to be annotated. Each individual 10 s segment was considered as an independent annotation task, provided on Amazon Mechanical Turk as one HIT (Human Intelligence Task). In order to prevent the same worker annotating overlapping segments, the data was split into batches containing segments located at least 15 sec-

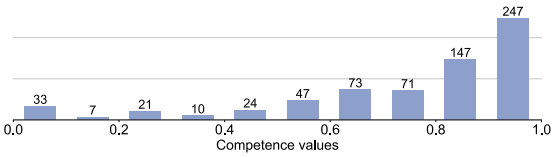

Figure 2: Annotator competence estimated using MACE.

onds apart in the original audio. The batches were then launched one at a time, and workers that already performed at least 50 hits in previous batch(es) were disqualified from working on the task. A payment of $0.10 was offered per HIT. Worker qualification was requested as at least 1000 completed HITs with average approval rating of at least 85%. One HIT consisted of listening to the provided audio excerpt and indicating which sounds are present in it, from the given list of six classes or "none of the above". The number of playbacks allowed was not limited. No visualization (e.g. spectrogram) was provided. Workers were instructed to complete the task using headphones, and in a quiet environment. Before the job, they were also provided short descriptions for every class, and four example sounds that contained events from all classes. The complete data annotation task was performed by 680 workers, with each 10 s segment being annotated by 5 workers. All jobs were accepted, in order to study the annotator behavior.

3.2. Analysis of annotation outcome

The correctness of the collected audio tags with respect to the generated reference labels was evaluated by considering the annotated segments individually. When the multiple opinions per segment are aggregated through majority vote, the comparison of the resulting tags with the ground truth achieves 68% F-score, with 98% precision, 52% recall. Using MACE to predict the true labels provides 86% F-score, with 97% precision and 77% recall, while union results in 78% F-score, with 70% precision and 89% recall. The relatively small recall indicates that many sounds are not annotated, possibly not being perceived in the audio mixture. With the majority vote, only slightly over half of the tags are found, while MACE raises the number of the tags correctly recalled to over three quarters. On the other hand, the very high precision indicates that the sounds which are labeled are labeled correctly. To keep this study focused on the annotation method itself, we do not analyze here the influence of SNR on recall. We note, however, that similar annotator behavior was observed as polyphony increases for annotators that selected onset and offset of sound instances [18].

The annotator competence analysis performed by MACE, presented in Fig. 2, reveals that 33 annotators have answered randomly, while 247 have a competence over 0.9. Inter-annotator agreement, measured using Krippendorff's alpha is 0.57. Eliminating annotators with low competence increases the agreement, e.g. a competence threshold of 0.6 increases α to 0.72, while a threshold of 0.8 increases α to 0.80. For further experiments, we use a competence threshold of 0.6, which keeps approximately 80% of annotators (538 of 680 annotators).

3.3. Estimation of strong labels

The temporal activity patterns of sound events is constructed as explained in Section 2, by stacking the annotated segments in their original order. In the basic setup, 5 opinions per 10 s segment result in 50 opinions for each 1 s of the estimated temporal activity

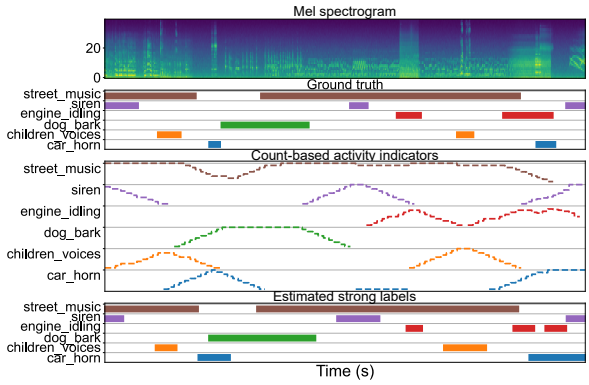

Figure 3: Estimation of strong labels based on the weak labels of consecutive, overlapping, audio segments.

(except the first and last 10 s of the original files). We estimate the temporal activity in three ways:

1. using data from all annotators, as explained above, with 50 opinions per 1 s segment.
2. using only annotators with competence higher than 0.6 (obtained using MACE). In this case, due to eliminating some of the annotations, the number of opinions per 1 s segment varies, being 37 on average
3. using the tags estimated using MACE as explained in Section 2. In this case there is one opinion per 10 s segment (the MACE output), resulting in 10 opinions per 1 s.

The resulting representation is then binarized using a threshold applied in each 1 s segment. Instead of the theoretical maximum value M for each 1 s, we use a threshold of 80%, to accommodate possible incorrect answers from the annotators. This reflects the proportion of annotators with an estimated competence over 0.6, as presented in Section 3.2. A sound event is therefore considered active in a 1 s segment if at least 80% of opinions available for that segment considered it active. Fig. 3 presents an example of this process, along with the ground truth for comparison. The quality of the resulting strong labels is evaluated by calculating detection metrics between them and the ground truth. For this, we calculate ER and F1 in 1 s segments (ER_1s, F1_1s), following the sound event evaluation procedure from DCASE Challenge [4], and the intersection-based F-score as defined in [19]. For the latter, we use two scenarios, as defined in DCASE 2021 Challenge Task 4[2], with different criteria for the detection tolerance criterion (DTC) and ground truth intersection criterion (GTC): DTC=GTC=0.7 (F1_dtc=0.7) and DTC=GTC=0.1 (F1_dtc=0.1). For details on the parameters and their effect, we refer the reader to [19]. The results are presented in Table 2.

A comparison of the estimated and the ground truth labels is presented in Fig. 4. This example shows that the proposed method has difficulty in estimating correctly the temporal activity for the short sound events. Even though most of the sounds in the example are identified, the high mismatch between the temporal boundaries of the short sounds will increase segment-based error rate and decrease F-score (as false positives or insufficient intersection). A more lenient intersection criterion (PSDS with DTC=0.1) results in an F-score of almost 90% for the best case.

[2]http://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments#evaluation

labels based on	ER_{1s}	S	D	I	$F1_{1s}$	P	R	$F1_{dtc=0.7}$	$F1_{dtc=0.1}$
all annotators	0.55	0.01	0.51	0.03	62.6	91.2	47.7	39.0%	67.4%
competence > 0.6	0.44	0.02	0.36	0.05	72.6	88.9	61.3	44.0%	81.3%
MACE	0.36	0.02	0.19	0.14	80.1	82.3	77.9	41.2%	89.5%

Table 2: Sound event detection scores between the estimated strong labels and the ground truth.

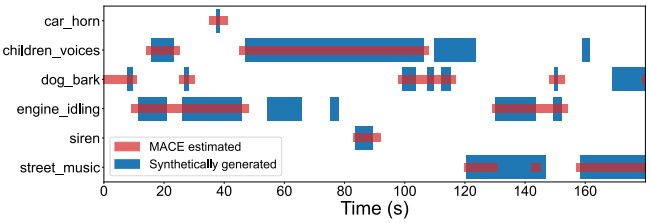

Figure 4: Visual comparison of generated and estimated labels

3.4. Discussion

A close analysis of the detection scores reveals that when relying on the majority vote among all annotators, the error rate is composed mostly of deletions, with only a small proportion of insertions and substitutions. This was expected based on the results from section 3.2, which indicated that many sounds were not annotated (recall 52%). MACE introduces many labels compared to the other methods, thus reducing deletions, but creates insertions because not all these labels are correct. The same trend can be seen in the dynamics shift of precision and recall: while the labels estimated based on all annotators have a high precision of 91.2%, but only 47.7% recall, MACE obtains a recall of 77.9% at the cost of reducing precision to 82.3%. Since tags produced by MACE for the 10 s segments had 86% F-score, with about one quarter of tags missing (recall 77%), a detection F-score of 89.5% (80.1% segment-based) between the estimated strong labels and ground truth ones means a very good match, provided that the missing tags may correspond to sound instances that were not perceived by the annotators.

4. SOUND EVENT DETECTION

As an additional experiment, we test how the mismatch between the estimated and synthetically generated strong labels affects SED evaluation. We use the system ranked best in the sound event detection task of DCASE 2017, where a similar amount of data was available. The system is a simple CRNN composed of 3 convolution blocks, each followed by batch normalization and max-pooling layers. The final two layers are composed of bi-directional gated recurrent units (GRU), in order to learn the temporal activity patterns. For more details of the model, we refer the reader to [20]. The system is trained using the audio data and the ground truth labels generated using Scaper. In order to use as much training data as possible, the train/test procedure follows a leave-one-out setup, in which one file is kept for testing, in turn, and the other 19 files are used for training and validation (18 for training, one for validation). The system output is evaluated against three different types of reference annotations:

1. Generated strong labels (ground truth)
2. Strong labels estimated using annotators with a competence higher than 0.6 (case 2 from Sec. 3.3)
3. Strong labels estimated using MACE (case 3 from Sec. 3.3)

eval. reference	ER_{1s}	$F1_{1s}$	$F1_{dtc=0.7}$	$F1_{dtc=0.1}$
GT	0.49	64.8%	26.4%	39.6%
estim.comp.>0.6	0.78	49.1%	12.5%	31.2%
estim. MACE	0.65	52.2%	13.0%	31.1%
train&eval MACE	0.58	55.9%	16.4%	40.3%

Table 3: Evaluation against different sets of strong labels

The results, evaluated using detection metrics, are presented in Table 3. We consider as a baseline performance the system trained and evaluated using the generated labels. Its error rate and F-score align well with the performance reported on other synthetic data, e.g. UrbanSED ($F1_{1s}$ approximately 60%) [16] and DCASE 2016 synthetic audio task (top systems had ER_{1s} 0.33-0.40 and $F1_{1s}$ 78-80%) [21]. When evaluated against the human-produced labels, the drop in measured performance is significant, even though what we evaluate is the exact same system output. If the same system is trained and evaluated using the estimated strong labels based on MACE, the measured performance is closer to the baseline performance (last row in Table 3). However, this system is trained and tested on approximately half the sound instances, as indicated by the low recall of the annotation process. These results show once more that the quality of annotations is a limiting factor not only in the training stage, but also for performance evaluation. The presented experiment is a typical situation, training SED systems on synthetic audio with correct and complete strong labels for training, and testing it on real-life recorded data. In addition to the presented effect of incomplete labels, testing on real recordings will introduce errors due to the mismatch in acoustic data. The presented system is a rather simple one, not considered state-of-the-art, therefore the effects of a weak system and the incomplete annotation are combined in the evaluated performance. However, we expect a similar effect of the annotations on more powerful systems too.

5. CONCLUSIONS

As sound event detection applications are moving towards systems applicable in real-life, a limiting factor of the development is the data annotation process. Even though training of systems can be achieved without strongly-labeled data, manually annotated data is necessary for evaluating the system behavior on real recordings corresponding to the user scenario. To alleviate the burden and subjectivity of manual annotation, we presented a method that can produce strong labels through crowdsourcing. Based on annotator competence estimation, a good, though incomplete, set of labels was produced. The resulting aggregated annotation is objective, being composed of multiple opinions. In future work, we will investigate further optimization of MACE for the case of strong labels, and investigate methods for producing the minimum amount of labels necessary for a reliable estimation, to reduce the redundancy of annotations where possible.

6. REFERENCES

[1] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[2] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019.

[3] A. Mesaros, T. Heittola, and D. Ellis, "Datasets and evaluation," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. Ellis, Eds. Springer, 2018, ch. 6, pp. 147–179.

[4] 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.

[5] 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[6] S. Hershey, D. P. W. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[7] C. Guastavino, "Everyday sound categorization," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. Ellis, Eds. Springer, 2018, ch. 7, pp. 183–213.

[8] T. Kauppi, J.-K. Kamarainen, L. Lensu, V. Kalesnykiene, I. Sorri, H. Kälviäinen, H. Uusitalo, and J. Pietilä, "Fusion of multiple expert annotations and overall score selection for medical image diagnosis," in *Image Analysis*, A.-B. Salberg, J. Y. Hardeberg, and R. Jenssen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 760–769.

[9] J.-K. Kamarainen, L. Lensu, and T. Kauppi, "Combining multiple image segmentations by maximizing expert agreement," in *Machine Learning in Medical Imaging*, F. Wang, D. Shen, P. Yan, and K. Suzuki, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 193–200.

[10] M. Cartwright, G. Dove, A. E. Méndez Méndez, J. P. Bello, and O. Nov, "Crowdsourcing multi-label audio annotation tasks with citizen scientists," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 111.

[11] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 21–25.

[12] E. Humphrey, S. Durand, and B. McFee, "OpenMIC-2018: An open data-set for multiple instrument recognition." in *ISMIR*, 2018.

[13] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning whom to trust with MACE," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 1120–1130.

[14] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds." *Journal of experimental psychology: human perception and performance*, vol. 19, no. 2, p. 250267, 1993.

[15] I. Martín-Morató and A. Mesaros, "What is the ground truth? reliability of multi-annotator data for audio tagging," in *29th European Signal Processing Conference 2019 (EUSIPCO 2019)*, 2021.

[16] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.

[17] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 10411044.

[18] M. Cartwright, A. Seals, J. Salamon, A. Williams, S. Mikloska, D. MacConnell, E. Law, J. P. Bello, and O. Nov, "Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–21, 2017.

[19] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 61–65.

[20] S. Adavanne, P. Pertil, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 771–775.

[21] 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.