PAPAFIL: A LOW COMPLEXITY SOUND EVENT LOCALIZATION AND DETECTION METHOD WITH PARAMETRIC PARTICLE FILTERING AND GRADIENT BOOSTING

Andrés Pérez-López[1,2]*, Rafael Ibáñez-Usach*[3]

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain, andres.perez@upf.edu
[2] Eurecat, Centre Tecnòlogic de Catalunya, Barcelona, Spain,
[3] STRATIO, Madrid, Spain, ribanez@stratio.com

ABSTRACT

The present article describes the architecture of a system submitted to the DCASE 2020 Challenge - Task 3: Sound Event Localization and Detection. The proposed method conforms a low complexity solution for the task. It is based on four building blocks: a spatial parametric analysis to find single-source spectrogram bins, a particle tracker to estimate trajectories and temporal activities, a spatial filter, and a single-class classifier implemented with a gradient boosting machine . Results from the development dataset show that the proposed method outperforms a deep learning baseline in three out of the four evaluation metrics considered in the challenge, and obtains an overall score almost ten points above the baseline.

Index Terms— SELD, ambisonics, tracking, event classification, gradient boosting

1. INTRODUCTION

Sound Event Localization and Detection (SELD) refers to the problem of identifying, for each individual event present in a sound field, the **spatial location** Ω, **temporal activity** Υ, and **sound class** κ to which it belongs.

The organization of a dedicated SELD task within the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 can be considered as a milestone for the development of the SELD research problem. Indeed, a large number of novel methodologies were developed for the Challenge, most of them based on Convolutional Recurrent Neural Networks (CRNN). The performance of the baseline method, a CRNN that performed jointly the localization and classification tasks [1], was vastly exceeded by a variety of deep-learning based algorithms [2, 3, 4]. Some of these improvements have been included in the baseline system for the SELD Challenge of DCASE 2020.

Despite the predominant trend towards high-complexity deep-learning architectures, some recent works have been able to match or even improve CRNN-based methods with regard to localization, by using parametric analysis of the ambisonic sound field [5, 6]. Apart from the benefit derived by their simplicity, these approaches are able to resolve the case of overlapping events of the same class, a situation difficult to disambiguate for CRNN-based methods [7].

The present work continues the exploration of possibilities of parametric SELD mehthods, focusing on a low-complexity architecture that makes use of traditional, feature-based machine learning techniques. The method been developed in the context of the SELD task withih DCASE 2020 Challenge, and therefore utilizes the proposed dataset, baseline system and evaluation metrics.

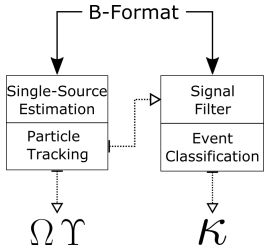

Figure 1: Architecture of the proposed methodology.

2. SYSTEM DESCRIPTION

The proposed method, referred to as *PAPAFIL*, can be summed up in four steps:

1. Estimate single-source time-frequency bins.
2. Use a particle tracking system to estimate event trajectories and activation times from single-source bins.
3. Perform spatio-temporal filtering on the input signal.
4. Assign a class label to the estimated event.

A scheme of the method is shown in Fig. 1. The full implementation is available online with an open source license [8].

2.1. Single-source estimation

The first step is the transformation of the B-Format input signal $\boldsymbol{x}(t) = [x_0(t), x_1(t), x_2(t), x_3(t)]^\intercal$ using the Short-Time Fourier Transform (STFT) into the time-frequency (TF) signal $\boldsymbol{X}(k, n)$, with k and n denoting the frequency and time indices, respectively. The number of channels of the input signal is $M = 4$.

In the resulting spectrogram, the frequencies above a given limit f_{max} are discarded; this procedure speeds up the method while maintaining the directional information, given that the microphone geometry aliases spatial measurements above approx. 5 kHz [9].

Assuming that the sources are sparse in time-frequency, it could be possible to identify TF bins which contain a significant energetic contribution from only one source. These bins could be then used to produce accurate Direction of Arrival (DOA) estimates. The effectiveness of this approach has already been demonstrated [10, 6].

Single-source TF bins are computed from the DirAC parametric analysis [11, 12] A variety of alternative subspace methods are known [13, 14]; however, those methods require local estimation of eigenvalues, which is a computationally expensive procedure.

A TF bin is counted as single-source if its diffuseness $\Psi(k, n)$ is lower than a threshold Ψ_{max}. Diffuseness is defined as [12]:

$$\Psi = 1 - 2\frac{\|\langle \Re\{X_0^*[X_1, X_2, X_3]\}\rangle\|}{\langle |X_0|^2 + \|[X_1, X_2, X_3]\|^2\rangle}, \quad (1)$$

where the time and frequency indices have been dropped for clarity, and $\langle\cdot\rangle$ represents the temporal expectation operator, which is usually implemented by averaging over N_Ψ neighbor frames.

Finally, the DOA $\Omega(k, n)$ of the TF bins passing the aforementioned single-source test is computed as the angle of the active intensity vector [12]:

$$\Omega = \angle(\Re\{X_0^*[X_1, X_2, X_3]\}), \quad (2)$$

where $\angle$ is the spherical angle operator. To illustrate the process, an example of the method output is plotted in Fig. 2 (top).

2.2. Particle tracking

Once a set of reliable TF DOA estimates is obtained, the next step is the generalization of the individual measurements into trajectories and temporal activations. In our case, we opted for the Rao-Blackwellized Monte-Carlo Data Association (RBMCDA) algorithm [15], which decomposes the multiple target tracking problem in two: it solves first the data association problem, and then performs the single target tracking individually. This method has been recently used in the context of sound event localization and tracking with successful results [1, 16]; the code used for our implementation has been adapted from the same authors [17].

The system takes as the input the set of TF DOA values passing the single source test, and produces spatio-temporal event trajectories, considering an event as an entity with contiguous temporal activation and continuous spatial position. More specifically, for each time frame of the DOA masked spectrogram, the median[1] of all narrowband DOA estimates is computed. The resulting value is added to the measurement space of the tracker if the number of single-source frequency bins for that frame exceeds a minimum K_{min}.

The performance of the RBMCDA algorithm is controlled by several parameters. Some of the most relevant include the angular velocity prior v, the standard deviation σ_ν and the spectral density s_ν of the measurement noise, the prior probabilities of birth p_{birth} and noise percentage p_ν, and the number of Monte-Carlo particles N. Position-related parameters are adjusted with respect to their ranges, so that azimuth-related magnitudes double elevation values.

The procedure is followed by a numerical post-processing step, which includes data interpolation, resampling (if needed), and removal of elements shorter than T_{min}. Finally, the system provides a list of J events, each one having an instantaneous position $\Omega_j(t)$ and a temporal activation Υ_j. An example of the system inputs and outputs is depicted in Fig. 2 (bottom).

2.3. Signal filter

The information provided by the particle tracking system is used to spatially filter the input signal. This can provide an enhanced monophonic estimate of an event $\tilde{s}_j(t)$ with reduced influence of simultaneous events. The process is performed by steering a virtual first-order cardioid in the direction of interest:

$$\tilde{s}_j(t) = \sum_{m=0}^{M-1} x_m(t) Y_m(\Omega_j)\alpha_n \quad (3)$$

[1]Circular median in the case of azimuth.

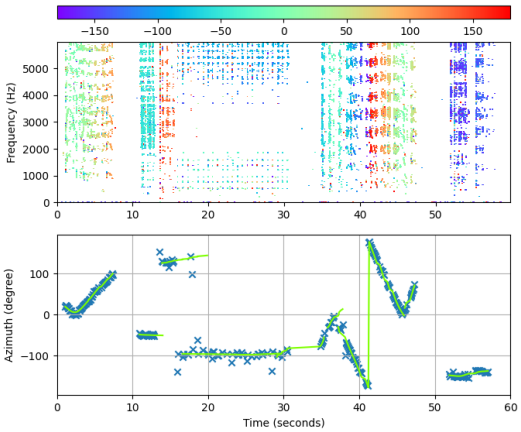

Figure 2: Estimation of localization and tracking. Top: azimuth spectrogram after diffuseness mask; color indicates estimated position (in degrees) of a TF bin passing the single-source test. Bottom: input/output of the particle tracking; the crosses represent the measurement space, and the continuous lines are the resulting events.

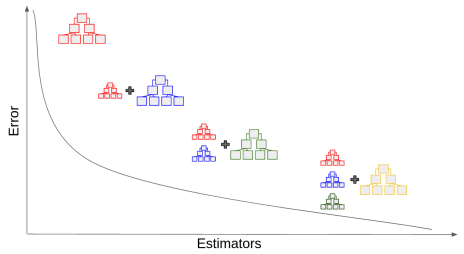

Figure 3: Gradient boosting machine learning process. Adding weak estimators allows reducing overall error in the predictions.

where $\boldsymbol{Y}(\Omega_j) = [Y_0(\Omega_j), \ldots, Y_3(\Omega_j)]^\intercal$ are the real-valued spherical harmonics up to first order evaluated in the direction Ω_j [18], and the column vector α_n controls the beam pattern directivity.

2.4. Event classification

As a final step, a class label is assigned to each estimated event $\tilde{s}_j(t)$ using a single-class classifier. Since the objective is to keep complexity low and make results interpretable, a machine learning algorithm is used instead of deep learning frameworks. The main advantages of this choice are: (i) low number of parameters; (ii) low train and predict computational time, easing reproducibility; and (iii) relative importance of the features in the output can be interpreted, which is not possible with deep learning approaches.

Gradient Boosting Machine (GBM, Fig. 3) has been selected as the classification algorithm since it is a powerful yet simple technique for predictive modelling. In essence, the algorithm is aimed to minimize the loss of the objective function by adding many weak learners. These learners are typically simple decision trees and their parameters are tuned using gradient descent techniques. GBM implementation makes use of the *scikit-learn* library [19].

Sound features are obtained using extractors from *Essentia*, an open-source library for audio analysis [20]. Given the heterogeneous nature of the sound classes, a mixture of spectral, temporal and harmonic features are used, as shown in Table 1. Features are computed either frame-based or on the whole event; in the former case, the classifier is fed with their temporal first-order statistics.

Table 1: Acoustic features used for classification, grouped by type.

Type	Features	Number
Low-level	Melbands	24
	MFCC	13
	Spectral Features	26
SFX	Duration	2
	Harmonic	4
	Sound envelope	11
	Pitch envelope	4

3. EXPERIMENTS

3.1. Dataset and baseline system

The dataset used is the FOA subset of the development set of the *TAU-NIGENS Spatial Sound Events 2020* [7], which features 600 different B-Format clips of 60 seconds long each. Each clip contains multiple sound events, which belong to one of the fourteen sound classes from the NIGENS database [21]. Events are also located at a potentially time-varying positions, and the maximum instantaneous overlapping of sources allowed is limited to two. Fifteen different Room Impulse Responses (RIR) are used for scene reverberation, covering a vast range of acoustic conditions. Furthermore, the audio clips contain a moderate amount of recorded background sounds.

The baseline method is based on the recently proposed SELDnet architecture [1], which features a 3-layer Convolutional Recurrent Neural Network (CRNN) that solves both localization and classification problems jointly. Additionally, the baseline implementation has been improved with several changes inspired by one of the best performing methods in DCASE 2019 Task 3 Challenge [3].

3.2. Experimental setup

In order to explore the performance of the system, two different approaches have been undertaken regarding the creation of the training dataset for the monophonic single-class classifier. The first approach, referred to as *PAPAFIL1*, collects all event localization, temporal activation and class information by parsing the annotation files. Conversely, the second approach, called *PAPAFIL2*, uses the proposed parametric particle filter to estimate localizations and activations, and the class label is assigned to each event by a custom association algorithm based on spatio-temporal distance. In both cases, the input signal is filtered with the obtained information in order to conform the monophonic event estimates.

Therefore, the difference between training datasets is noticeable: while the training events in *PAPAFIL1* are more accurately determined than in *PAPAFIL2*, the differences with respect to the prediction scenario are much bigger in the former case. The number of individual events for each of the approaches is plotted in Fig. 4. Approximately half of the classes have similar number of instances in both datasets. However, the other half presents noticeable differences, which might be explained by the different criteria applied for the consideration of event temporal activations: the groundtruth seems to follow a frame-based activity detection approach, while the output of the proposed method tends to consider events as time-continuous manifestations, influenced by the particle filter.

This situation leads to two different *oracle* systems (referred to by appending *-O* in the method name), which represent the best performance theoretically achievable for the corresponding method.

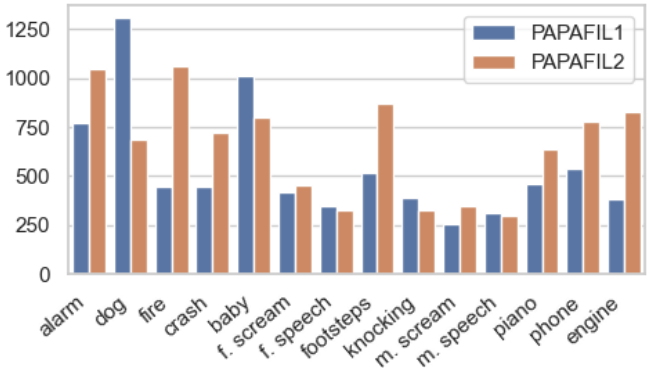

Figure 4: Number of occurrences of each event class in the training set, for both proposed methods.

The accurate information of the *PAPAFIL1* training set suggests a need for data augmentation; in contrast, the training material used in *PAPAFIL2* is already provided by a certain extent of variability. This situation motivates the implementation of data augmentation methods in the *PAPAFIL1* training set. Specifically, several standard data augmentation techniques are implemented: pitch shifting, time shifting, time stretching and white noise addition. Furthermore, given the observed high influence of reverberation in the system performance, a reverberant data augmentation technique based on synthetic RIRs has been considered. Ten different single-channel RIRs, with reverberation times between 0.3 and 1.1 seconds, have been synthetically created using the *masp* library [22]. During training, each event estimate is convolved with one of the RIRs, randomly chosen. RIR augmentation has recently been shown very effective for blind reverberation time estimation [23] but, to the best of the authors' knowlegdge, this is the first application in SELD.

Table 2 shows a comprehensive list of the parameters used throughout the different steps of the proposed method. All values are equal for both presented approaches, except for the number of Monte-Carlo particles N. The values for Single-Source Estimation and Particle Filtering parameters have been iteratively refined by manual tuning and inspection, departing from standard values. The beamforming weights α_m correspond to the *maximum directivity beamformer*, which minimizes the energy contributions from directions other than the lookup direction [24]. In the spatial audio field, such property is also known as the *max-rE* decoder [18]. Regarding event classification, a cross-validation scheme has been implemented for tuning GBM hyperparameters.

3.3. Evaluation metrics

The system is evaluated according to the joint metrics proposed in the Challenge [25]. The metrics evaluate jointly the localization and the classification, and are divided into two types: location-aware classification, and classification-aware localization. There are two classification metrics: Error Rate (ER_{20}) and F-Score (F_{20}). As the name suggests, the metrics are conditioned to a minimum localization performance, which is set to 20° in this case. Localization metrics are also two-fold: Localization Error (LE_{CD}) and Localization Recall (LR_{CD}); as their name suggests, the metrics are class-dependent, and thus are conditioned to a correct classification. Finally, the SELD score is an average of the four other metrics, used to conveniently sum up the results.

Table 2: (Hyper-)parameter values.

Step	Parameter	Value	Unit
Single-Source	sample rate	24	kHz
	window size	2400	samples
	window overlap	50	%
	f_{max}	6	kHz
	N_Ψ	2	frames
	Ψ_{max}	0.1	
Particle Filtering	v	2	°/frame
	σ_ν	5	
	s_ν	20	
	p_{birth}	0.25	
	p_ν	0.25	
	N	100 / 30	
	K_{min}	10	bins/frame
	T_{min}	10	frames
Signal Filter	α_0	0.775	
	α_1	3 * 0.4	
Event Classification	number of estimators	1300	trees
	loss	*mlogloss*	
	learning rate	0.05	
	max depth	4	
	min samples leaf	10	samples

4. RESULTS

Table 3 summarizes the results of the experiments using the recommended data split: training with folds 3 to 6, validation with fold 2 and testing with fold 1. Results are reported for three different systems: the baseline and the two proposed methods *PAPAFIL1* and *PAPAFIL2*. The results of their respective oracle results, *PAPAFIL1-O* and *PAPAFIL2-O*, are also provided.

Both proposed approaches outperform the baseline system in three out of the four evaluation metrics (ER_{20}, F_{20} and LE_{CD}). Although the results obtained by both of them are similar, *PAPAFIL2* obtains better classification scores (ER_{20} and F_{20}), and *PAPAFIL1* performs subtly better regarding localization error (LE_{CD}). However, the localization recall results (LR_{CD}) are slightly worst than the baseline in both cases. This fact does not prevent the proposed methods to have a SELD score better than the baseline: 0.41 (*PAPAFIL1*) and 0.38 (*PAPAFIL2*), against 0.47 (*BASELINE*).

The results obtained by the oracle methods are within the expected ranges. *PAPAFIL1-O* performs almost perfectly regarding LE_{CD}, but the classification errors influence the LR_{CD} result. In turn, *PAPAFIL2-O* performs better than *PAPAFIL1-O* regarding all metrics, excepting LE_{CD}; this improvement is specially noticeable in LR_{CD}, with a performance difference of about 15%. The good results obtained by *PAPAFIL2-O* validate the proposed particle filtering approach, an leave space for improvements that might be given by a better understanding and fine tuning of the model.

The performance of the proposed methods deteriorates noticeably with overlapping sounds. A closer inspection reveals that, in many occasions, the TF bins passing the single-source test mostly belong to one out of two simultaneous sources. It is a known issue that performance of DirAC diffuseness is reduced when two sources are present [13]; similar problems have been reported in [16], where an instantaneous source number estimator is used in combination with the particle filter. As in that case, the results suggest the need for more sophisticated source detection and counting methods.

Table 3: Evaluation results on the development set.

Method	ER_{20}	F_{20}	LE_{CD}	LR_{CD}	SELD
BASELINE	0.72	37.4 %	22.8°	**60.7** %	0.47
PAPAFIL1	0.60	49.8 %	**13.4**°	54.4 %	0.41
PAPAFIL2	**0.57**	**54.0** %	13.8°	59.7 %	**0.38**
PAPAFIL1-O	0.37	67.0 %	2.0°	68.6 %	0.26
PAPAFIL2-O	0.32	79.6 %	8.5°	82.4%	0.19

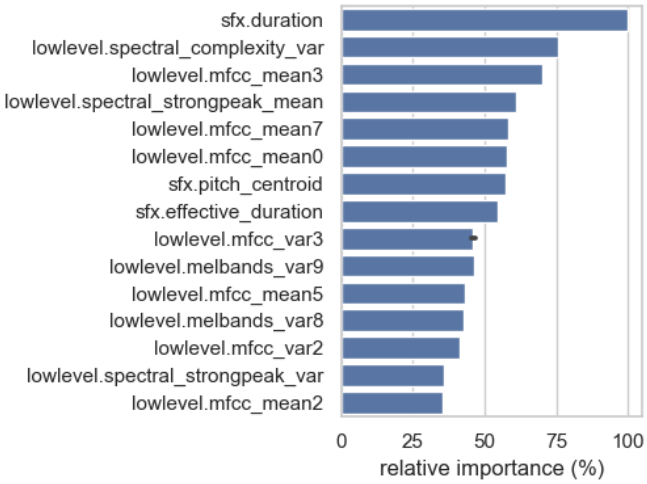

Figure 5: Most representative features in event classifier.

To conclude the analysis, Fig. 5 shows the relative importance of the fifteen most relevant acoustic features for the *PAPAFIL2* classifier model. Event duration is clearly the feature with the highest importance, and effective duration (duration of the signal discarding silence) also appears in the eighth position. This fact can help to explain the better performance of *PAPAFIL2* over *PAPAFIL1*: the temporal activities of the events in training and prediction are much more similar to each other in the former method, as a consequence of the training set generation approach. Furthermore, it is interesting to notice the high relevance of low-level features, and specifically several MFCC combinations (eight of the fifteen reported features) and various extractors related to the spectral structure. The absence of pitch, harmonic and envelope features in the list represents a significant finding as well.

5. CONCLUSION

We present a novel low-complexity method for Sound Event Localization and Detection of First Order Ambisonic signals, based on four steps: estimation of single-source spectrogram regions by parametric analysis; computation of event trajectories and activations by means of a particle tracker; spatio-temporal filtering of the input signal; and single-class monophonic event classification by Gradient Boosting. Results show that the proposed method outperforms the baseline method, a state-of-the-art Convolutional Recurrent Neural Network. Specifically, our method is able to improve the baseline SELD score by almost ten points, while increasing the scores in three out of the four metrics under consideration.

6. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, March 2018.

[2] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," *arXiv preprint arXiv:1908.00766*, 2019.

[3] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 30–34.

[4] F. Grondin, F. Glass, I. Sobieraj, and M. D. Plumbley, "Sound event localization and detection using crnn on pairs of microphones," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 84–88.

[5] A. Pérez-López, E. Fonseca, and X. Serra, "A hybrid parametric-deep learning approach for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 189–193.

[6] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "A sequence matching network for polyphonic sound event localization and detection," in *Proc. IEEE ICASSP*. IEEE, May 2020, pp. 71–75.

[7] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv e-prints: 2006.01919*, 2020.

[8] https://github.com/andresperezlopez/DCASE2020.

[9] S. Bertet, J. Daniel, and S. Moreau, "3D Sound Field Recording With Higher Order Ambisonics - Objective Measurements and Validation of a 4th Order Spherical Microphone," in *Proc. 120th AES Convention*, Paris, France, May 2006, pp. 1–24.

[10] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust doa estimation of multiple speech sources," in *Proc. IEEE ICASSP*. Florence, Italy: IEEE, May 2014, pp. 2287–2291.

[11] J. Merimaa and V. Pulkki, "Spatial impulse response rendering i: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, December 2005.

[12] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, June 2007.

[13] N. Epain and C. T. Jin, "Spherical harmonic signal covariance and sound field diffuseness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1796–1807, June 2016.

[14] L. Madmoni and B. Rafaely, "Direction of arrival estimation for reverberant speech based on enhanced decomposition of the direct sound," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 131–142, 2018.

[15] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-blackwellized monte carlo data association for multiple target tracking," in *Proceedings of the Seventh International Conference on Information Fusion*, vol. 1, Stockholm, Sweden, June 2004, pp. 583–590.

[16] S. Adavanne, A. Politis, and T. Virtanen, "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," *arXiv preprint arXiv:1904.12769*, 2019.

[17] https://github.com/sharathadavanne/multiple-target-tracking.

[18] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, University of Paris VI, 2000.

[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, November 2011.

[20] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer, O. Mayor, G. Roma Trepat, J. Salamon, J. R. Zapata González, X. Serra, *et al.*, "Essentia: An audio analysis library for music information retrieval," in *14th Conference of the International Society for Music Information Retrieval (ISMIR); p. 493-8.*, Curitiba, Brazil, November 2013.

[21] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The nigens general sound events database," *arXiv preprint arXiv:1902.08314*, 2019.

[22] A. Pérez-López and A. Politis, "A python library for multichannel acoustic signal processing," in *Proc. 148th Audio Engineering Society Convention*. Audio Engineering Society, May 2020.

[23] N. J. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," in *Proc. IEEE ICASSP*. IEEE, 2020.

[24] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

[25] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2019.