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

Technical Report

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 technical report describes the architecture of the 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 gradient boosting machine single-class classifier. Provisional results, computed from the development dataset, show that the proposed method outperforms a CRNN 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. SIGNAL MODEL

The input signals under consideration follow the convention used by the TAU-NIGENS Spatial Sound Events 2020 - FOA dataset [1]. Each audio file has a duration of 60 seconds, and presents a First Order Ambisonics (FOA) signal, following ACN and SN3D conventions [2]. Each FOA clip contains the spatial representation of a a reverberant sound scene, composed of an arbitrary number of individual sound events plus background noise.

The individual sound events are taken from the NIGENS, which features over a thousand instances belonging to 14 different sound classes [3]. Events have been reverberated using real ambisonics Room Impulse Responses (RIRs). Furthermore, the sound events are distributed over the space, and can be either static or dynamic; in the latter case, the movements are always circular around the listener. Only a maximum of two events can be instantaneously active.

We can use the following model to describe the audio scenes:

$$\boldsymbol{x}(t) = \sum_{j=1}^{J} s_j^{\kappa}(t-\tau_j) * \boldsymbol{h}(t,\Omega_j) + \boldsymbol{\nu}(t), \qquad (1)$$

where $\boldsymbol{x}(t) = [x_0(t), x_1(t), x_2(t), x_3(t)]^{\intercal}$ represent the FOA (with $M = 4$ channels) recorded signal, composed of J different sound events, $s_j^{\kappa}(t), j = 1, \dots, J$, each one belonging to a different class κ. Each event is convolved with an ambisonic room impulse response $\boldsymbol{h}(t,\Omega)$, which encodes the (potentially time-dependent) position of the source $\Omega = (\varphi, \theta)$ as the azimuth and elevation angles in spherical coordinates, respectively. Furthermore, each event is delayed an arbitrary amount of time τ, and has a duration T. The temporal information can be summed up as the activation $\Upsilon = (\tau, \tau + T)$, which contains the *onset* and *offset* times of the event. Finally, the term $\boldsymbol{\nu}(t)$ models the background noise present in the sound scene.

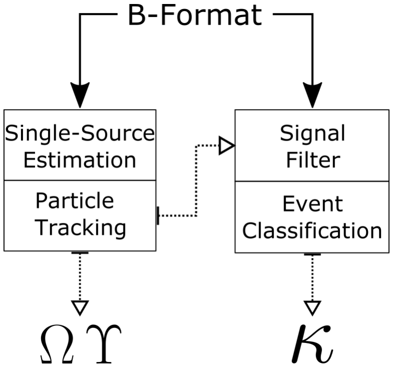

Figure 1: Architecture of the proposed methodology.

According to (1), given the observed reverberant FOA signal $\boldsymbol{x}(t)$, the Sound Event Localization and Detection (SELD) problem consists in the estimation of the following parameters for each sound event $s_j^{\kappa}(t)$: the **instantaneous localization** Ω, the **activation times** Υ and the **class** κ to which it belongs.

2. SYSTEM DESCRIPTION

The proposed method can be summed up in four steps:

1. Estimate single-source time-frequency bins from the input signal.
2. Use a particle tracking system to convert then into event trajectories and activation times.
3. Perform spatio-temporal filtering on the input signal with the resulting estimations.
4. The spatially filtered signal is assigned a class label by a classifier.

A scheme of the method is shown in Figure 1. A full implementation of the system can be found at `https://github.com/andresperezlopez/DCASE2020` with an open-source license.

2.1. Single-source estimation

The first step is the transformation of the input signal $\boldsymbol{x}(t)$ 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 frequencies of the resulting spectrogram above a given limit f_{max} are discarded; this procedure helps to speed up the process while maintaining most of the directional information, given that the microphone geometry (with radius $R = 0.042$ m) provides aliased spatial measurements above 5 kHz approximately [4].

Assuming that the sources have a sparse TF representation, it could be possible to identify which TF bins contain a significant energetic contribution from one only source, i.e., without significant cross-talk from other sources or background noise. These TF bins could be then used to compute accurate Direction of Arrival (DOA) estimates.

We compute the single-source TF bins from the DirAC parametric analysis [5, 6] A variety of alternative methods are known, mostly based on subspace analysis [7, 8]; however, those methods require local estimation of eigenvalues, which is a computationally complex procedure. This is the main reason for the choice of DirAC-based analysis in this work.

A TF bin is considered to be single-source if its diffuseness $\Psi(k,n)$ is lower than a given threshold Ψ_{min}. Diffuseness is computed as [6]:

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

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, we compute the DOA $\Omega(k,n)$ of the TF bins passing the single-source test. DOA is computed as the angle of the active intensity vector [6]:

$$\Omega = \angle(\Re\{X_0^*[X_1, X_2, X_3]\}), \tag{3}$$

where $\angle$ is the spherical angle operator. Figure 3 (top) shows the estimated DOAs of the single-source TF bins, for an example signal input.

2.2. Particle tracking

Once a set of reliable 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 [9], which decomposes the 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 [10, 11]; the code used here has been adapted from these authors[1].

The system takes as the input the set of TF DOA values passing the single source test, and produces spatio-temporal particle trajectories. More specifically, for each time frame of the DOA masked spectrogram, the median[2] of all frequency-dependent DOA estimates is computed. The resulting value will be added to the measurement space of the tracker if the number of frequencies passing the test exceeds a given minimum K_{min}.

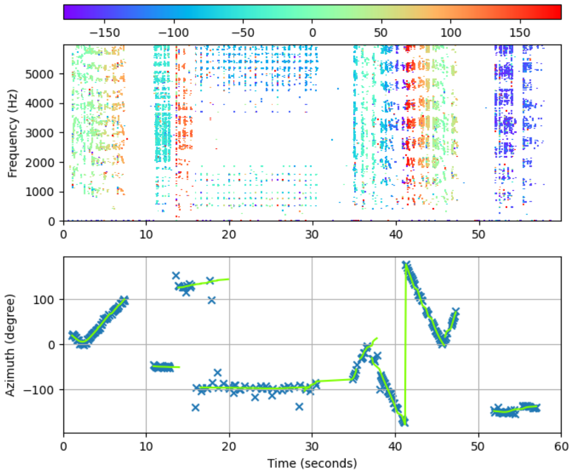

Figure 2: Estimation of localization and temporal activation. Top: azimuth spectrogram after diffuseness mask; color indicates estimated position of a TF bin passing the single-source test. Bottom: input/output of the particle tracking; the crosses represent the measurement space, and the continuous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

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. All parameters related to position are adjusted with respect to their ranges, so that the azimuth magnitude is twice as big as the elevation magnitude. A more detailed insight on the method is outside the scope of the document.

The procedure is followed by a numerical post-processing step, which includes data interpolation, resampling (if the processing was performed using a frame size different than the target of 0.1 s) and removal of elements shorter than T_{min}. Finally, the system provides a list of J events, each one having an instantaneous position Ω_j and a temporal activation $\Upsilon)j$, as required for the challenge task. An example of the system inputs and outputs is depicted in Figure 3 (bottom).

2.3. Signal filter

The spatial estimates for each event, Ω_j, provided by the particle tracking system can be used to spatially filter the audio. This process is performed by steering a virtual first-order cardioid in the direction of interest, by means of a linear combination of the input channels:

$$\tilde{s}_j(t) = \sum_{m=0}^{M-1} x_m(t) Y_m(\Omega_j)\alpha_m \tag{4}$$

where $\boldsymbol{Y}(\Omega_j) = [Y_0(\Omega_j), \ldots, Y_{M-1}(\Omega_j)]^\intercal$ are the real-valued spherical harmonics up to first order evaluated in the direction Ω_j [12], and the column vector α controls the beam pattern directivity. The result of this process is a monophonic estimate for each event, $\tilde{s}_j(t)$, which is temporally delimited by Υ_j. As a last

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

[2] Circular median in the case of azimuth.

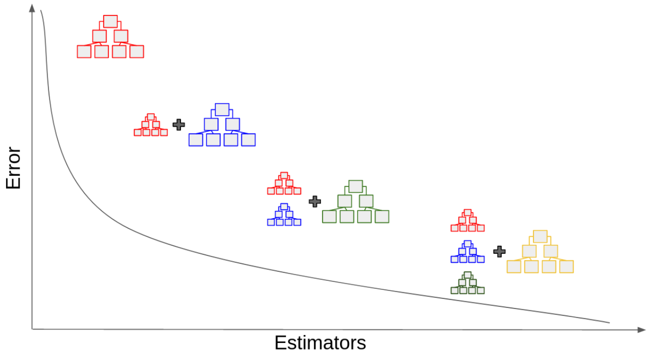

Figure 3: Gradient boosting machine learning process. Adding weak estimators allows reducing overall error in the predictions.

step, each estimate is amplitude peak-normalized, in order to minimize potential amplitude variability due to arbitrary configurations of the sound scene.

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:

- Low number of parameters.
- Low train and predict computational time, allowing easy replication of the results.
- Relative importance of the features in the output can be interpreted which is not possible with deep learning approaches.

Gradient boosting machine (GBM) 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. Specifically, XGBoost framework is implemented for training process due to its proven performance in a wide range of classification problems [13].

Sound features are obtained using extractors from Essentia, an open-source library for audio analysis [14]. Given the heterogeneous nature of the sound classes included in the dataset, a mixture of spectral, temporal and harmonic feature extractors have been selected, as shown in Table 1.

3. EXPERIMENTS

3.1. Dataset and baseline system

The baseline method is based on the recently proposed SELDnet architecture [10], which features a Convolutional Recurrent Neural Network (CRNN) that solves both localization and classification problems jointly. Additionally, the SELDnet implementation for the challenge has been improved with several changes inspired by one of the best performing methods in last year's challenge [15].

code from libraries

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

Type	Features	Number
Low level	Melbands	24
	MFCC	13
	Spectral Features	25
	Pitch Salience	1
SFX	Total and perceived sound duration	2
	Descriptors based on pitch and harmonics estimation	4
	Sound envelope descriptors	11
	Pitch envelope descriptors	4

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, and uses this oracle information to spatially filter the signal and label the monophonic estimates with which to train the classifier. 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 similarity. As in the previous case, the input signal is finally filtered to obtain the monophonic event estimates.

Therefore, there is a noticeable difference on the monophonic event datasets. While training events in *PAPAFIL1* are more accurately localized and detected than in *PAPAFIL2*, the differences with respect to the *evaluation* situation are much bigger in the former case. Consequently, a slightly better performance of the second method might be expected, provided that the parametric particle filter performance has some degree of robustness and accuracy.

The perfect localization and temporal activity information in the *PAPAFIL1* training set suggests a need for data augmentation techniques. In contrast, the training material used in *PAPAFIL2* is already provided of a certain extent of variability. This situation motivates the implementation of data augmentation methods only 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 high observed influence of the reverberation in the system performance, a for reverberant data augmentation based on synthetic RIRs has been considered. This approach has recently been shown very effective for the blind reverberation time estimation [16] but, to the best of the authors' knowlegdge, this is the first application to the SELD problem. Regarding the specific implementation, ten different single-channel RIRs, with reverberation times in the range of 0.3 to 1.1 seconds, have been synthetically created by the Image Source method [17]. During training, each event estimate is convolved with one of the RIRs, randomly chosen.

The presented scheme leads to two different *oracle* systems (named by an *-O* appendix in the method name), which represent the best performance theoretically achievable for the correspond-

Table 2: (Hyper-)parameter values.

Step	Parameter	Value	Unit
Single-Source Estimation	sample rate	24	kHz
	window type	hann	
	window size	2400	samples
	window overlap	50	%
	f_{max}	6	kHz
	N_Ψ	2	frames
	Ψ_{min}	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	
	α_2	3 * 0.4	
	α_3	3 * 0.4	
Event Classification	number of estimators	1300	trees
	loss	mlogloss	
	learningrate	0.05	
	maxdepth	4	
	minsamplesleaf	10	samples

ing method. In this way, it is expected that *PAPAFIL1-O* obtains very high localization scores overall, while the performance of *PAPAFIL2-O* can be much similar to the non-oracle case.

It is important to notice that the spatial filtering is performed with a first-order cardioid, which provides a broad directive pattern. Accordingly, in the case of overlapping events, there will be always signal cross-talk, even when using the groundtruth annotations. The usage of higher ambisonic orders could easily mitigate this effect.

Table 2 shows a comprehensive list of the parameters used throughout the different steps of the proposed method. All values are the same 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, starting 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 [18]. In the spatial audio field, such property is also known as the *max-rE* decoder [12]. 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 presented in [19]. 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°. Localization metrics are also twofold: Localization Error (LE$_{CD}$) and Localization Recall (LR$_{CD}$). Again, the subscript *CD* stands for *class-dependent*; thereby a correct localization evaluation is conditioned to a correct classification.

4. RESULTS

Table 3 summarizes the results of the experiments, according to two different evaluation setups. The top sub-table present the results using the following data split: training with folds 3 to 6, validation with fold 2 and testing with fold 1. This structure has been promoted by the Challenge organization as a fair way of comparing methods; furthermore, it is expected to provide similar results to the evaluation set, given that part of the data remains unseen at training. Conversely, the lower half of the table gives the results for the entire development dataset.

The table reports the results for three different systems. The baseline method has results reported only for the comparison split, while the proposed methods *PAPAFIL1* and *PAPAFIL2* and their respective oracle results *PAPAFIL1-O* and *PAPAFIL2-O* are reported for both sets.

Regarding the comparison split, the proposed methods outperform the baseline system in three out of the four evaluation metrics: ER$_{20}$, F$_{20}$ and LE$_{CD}$. While the results obtained by both of them are close, *PAPAFIL2* obtains better classification scores (ER$_{20}$ and F$_{20}$), and *PAPAFIL1* obtains subtly better localization error (LE$_{CD}$). However, the localization recall results (LR$_{CD}$) are slightly worst than the baseline. This fact does not prevent the proposed methods to have an overall score (SELD) better than the baseline: 0.41 (1) and 0.38 (2), agains 0.47.

With respect to the results evaluated on the full dataset, comparison with the baseline is not possible due to lack of results. However, when comparing the proposed approaches, *PAPAFIL2* outperforms in all evaluation metrics, scoring over ten points better in all metrics (including SELD) excepting $textLE_{CD}$, where the improvement is more moderate.

The results obtained by the oracle methods are within the expected ranges. *PAPAFIL1-O* performs almost perfectly on the entire dataset, and decreases its classification scores for the unseen split set; localization remains highly accurate, but the classification errors influence the $textLE_{CD}$ result. The performance of *PAPAFIL2-O* does not vary significantly when comparing evaluation sets, since it does not depend on training, and the differences are solely due to differing acoustic properties. On the other hand, the difference with respect to the non-oracle version is noticeably low in the case of the unseen split, with a result only slightly better: 0.24 (*PAPAFIL2-O*) vs 0.26 (*PAPAFIL2*).

Although the localization performance is good in general terms with both *PAPAFIL* approaches, the results deteriorate noticeably with overlapping sound, as the comparatively low score for the Localization Recall LR$_{CD}$ reflects. 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 DirAC diffuseness performance is reduced when two sources are present [7]. However, related problems have been observed in [11], where an instantaneous source number estimator is used together with the particle filter. In a similar manner as reported in that work, the results here suggest the need need for more sophisticated source detection and counting methods.

Table 3: Evaluation results on the development set. Top: results from the recommended comparison split. Bottom: overall results

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
PAPAFIL1-O	0.37	67.0 %	2.0°	68.6 %	0.26
PAPAFIL2	**0.57**	**54.0** %	13.8°	59.7 %	**0.38**
PAPAFIL2-O	0.32	79.6 %	8.5°	82.4%	0.19
PAPAFIL1	0.57	55.6 %	15.6°	66.7 %	0.36
PAPAFIL1-O	0.08	93.7 %	0.2°	94.0 %	0.05
PAPAFIL2	**0.44**	**68.0** %	**13.3**°	**79.6** %	**0.26**
PAPAFIL2-O	0.41	71.1 %	12.3°	82.0%	0.24

5. CONCLUSION

We have presented a novel method for Sound Event Localization and Detection (SELD), based on parametric particle filtering and gradient boosting single-class event classification of audio features. Results show that the proposed method outperforms the baseline method, a state-of-the-art Convolutional Recurrent Neural Network (CRNN). Specifically, the proposed method is able to improve the baseline SELD score by almost ten points, and to improve also in three out of the four proposed evaluation metrics, by means of a low complexity machine learning architecture.

6. REFERENCES

[1] 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. [Online]. Available: https://arxiv.org/abs/2006.01919

[2] T. Carpentier, “Normalization schemes in ambisonic: does it matter?” in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

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

[4] 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,” *120th AES Convention*, pp. 1–24, 2006.

[5] 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, 2005.

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

[7] 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, 2016.

[8] 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.

[9] 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. I, 2004, pp. 583–590.

[10] 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. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8567942

[11] 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.

[12] 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.

[13] T. Chen and C. Guestrin, “Xgboost, a scalable tree boosting system.”

[14] 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 *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil. ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR), 2013.

[15] 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.

[16] N. J. Bryan, “Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1–5.

[17] A. Perez-Lopez and A. Politis, “A python library for multichannel acoustic signal processing,” in *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020.

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

[19] 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, Oct 2019, accepted.