

DIRECTION OF ARRIVAL ESTIMATION FOR MULTIPLE SOUND SOURCES USING CONVOLUTIONAL RECURRENT NEURAL NETWORK

Sharath Adavanne^{*,1}, Archontis Politis^{*,2} and Tuomas Virtanen¹

¹ Laboratory of Signal Processing, Tampere University of Technology, Finland

² Department of Signal Processing and Acoustics, Aalto University, Finland

ABSTRACT

This paper proposes a deep neural network for estimating the directions of arrival (DOA) of multiple sound sources. The proposed stacked convolutional and recurrent neural network (DOAnet) generates a spatial pseudo-spectrum along with the DOA estimates in both azimuth and elevation. We avoid any explicit feature extraction step by using the magnitude and phase of the spectrogram as input to the network. The proposed DOAnet is evaluated by estimating the DOAs of multiple concurrently present sources in anechoic, matched and unmatched reverberant conditions. The results show that the proposed DOAnet is capable of estimating the number of sources and their respective DOAs with good precision and generate SPS with high signal-to-noise ratio.

Index Terms— Direction of arrival estimation, multiple sources, Ambisonic, convolutional recurrent neural network

1. INTRODUCTION

Direction of arrival (DOA) estimation is the task of identifying the relative position of the sound sources with respect to the microphone. DOA estimation is a fundamental operation in microphone array processing and forms an integral part of speech enhancement [1], distant automatic speech recognition [2], spatial audio coding [3], and sound source separation [4]. Popular approaches to DOA estimation are based on time-delay-of-arrival (TDOA) [5], the steered-response-power (SRP) [6], or on subspace methods such as multiple signal classification (MUSIC) [7] and the estimation of signal parameters via rotational invariance technique (ESPRIT) [8].

The aforementioned methods differ from each other in terms of algorithmic complexity, and their suitability to various arrays and sound scenarios. MUSIC specifically is very generic with regards to array geometry, directional properties and can handle multiple simultaneously active narrowband sources. On the other hand, MUSIC and subspace methods in general, require a good estimate of the number of active sources, which are often unavailable or difficult to obtain.

Furthermore, MUSIC is computationally intensive and can suffer at low signal to noise ratio (SNR) and in reverberant scenarios [9]. In this paper, we propose to overcome the above shortcomings with a deep neural network (DNN) method, referred to as DOAnet, that learns the number of sources from the input data, generates high precision DOA estimates and is robust to reverberation.

Recently, several DNN-based approaches have been proposed for DOA estimation [10–15]. There are six significant differences between them and the proposed method: a) All the aforementioned works focused on azimuth estimation, with the exception of [14] where the 2-D Cartesian coordinates of sound sources in a room were predicted, and [10] trained separate networks for azimuth and elevation estimation. In contrast, we demonstrate the estimation of both azimuth and elevation for the DOA by sampling the unit sphere uniformly and predicting the probability of sound source at each direction. b) The past works focused on the estimation of a single DOA at every time frame, with the exception of [12] where localization of azimuth for up to two sources simultaneously was proposed. On the other hand, DOAnet does not explicitly limit the number of directions to be estimated.

c) Past works were evaluated with different array geometries making comparison difficult. Although the DOAnet can be applied to any array geometry, we evaluate the method using real spherical harmonic input signals, which is an emerging popular spatial audio format under the name Ambisonics. Microphone signals from various arrays, such as spherical, circular, planar or volumetric, can be transformed to Ambisonic signals by an appropriate transform [16], resulting in a common representation of the 3-D sound recording. Although the DOAnet is scalable to higher-order Ambisonics, in this paper we evaluate it using the compact four-channel first-order Ambisonics (FOA).

d) Regarding classifiers, earlier methods have used fully connected (FC) neural networks [10–14] and convolutional neural networks (CNN) [15]. In this work, along with the CNNs we use recurrent neural network (RNN) layers. The usage of RNN allows the network to learn long-term temporal information. Such an architecture is referred to as a convolutional recurrent neural network (CRNN) in literature and is the state-of-the-art method in many single- [17, 18] and mul-

^{*}Equally contributing authors in this paper. The research leading to these results has received funding from the European Research Council under the European Unions H2020 Framework Programme through ERC Grant Agreement 637422 EVERYSOUND. The authors also wish to acknowledge CSC-IT Center for Science, Finland, for computational resources

tichannel [2, 19] audio tasks. e) Previous methods used inter-channel features such as generalized cross-correlation with phase transform (GCC-PHAT) [11, 14], eigen-decomposition of the spatial covariance matrix [12], inter-channel time delay (ITD) and inter-channel level differences (ILD) [10, 13]. More recently, Chakrabarty et al. [15] proposed to use only the phase component of the spectrogram, avoiding explicit feature extraction. In the proposed method, we use both the magnitude and the phase component. Contrary to [15], which employed omnidirectional sensors only, general arrays with directional microphones additionally encode the DOA information in magnitude differences, while Ambisonics especially encode directional information mainly in the magnitude component. f) All previous methods were evaluated on speech recordings that were synthetically spatialized and spatially static. We continue to use the static sound sources in the present work and extend them to a larger variety of sound events, such as impulsive and transient sounds.

2. METHOD

The block diagram of the proposed DOAnet is presented in Figure 1. The DOAnet takes multichannel audio as the input and first extracts the spectrogram. The phase and the magnitude component of this spectrogram are mapped using a CRNN to two outputs sequentially. The first output, spatial pseudo-spectrum (SPS) is generated as a regression task, followed by the DOA estimates as a classification task. The DOA is defined by the azimuth ϕ and elevation λ with respect to the microphone and the SPS is the intensity of sound along the DOA given by $S(\phi, \lambda)$.

In this paper, we use discrete ϕ and λ by uniformly sampling the 2-D polar coordinate space, with a resolution of 10 degrees in both azimuth and elevation, resulting in 614 sampled directions. The SPS is computed at each sampled direction, whereas, a subset of 432 directions is used for DOA, where the elevations are limited between -60 and 60 degrees.

2.1. Feature extraction

The spectrogram is calculated for each of the audio channels sampled at 44100 kHz. A 2048-point discrete Fourier transform (DFT) is calculated on Hamming windows of 40 ms with 50 % overlap. We keep 1024 values of the DFT corresponding to the positive frequencies, without the zeroth bin. L frames of features, each containing 1024 magnitude and phase values of the DFT extracted in all the C channels, are stacked in a $L \times 1024 \times 2C$ 3-D matrix and used as the input to the proposed neural network. The $2C$ dimension results from ordering the magnitude component of all channels first, followed by the phase. We use a sequence length L of 100 (= 2 s) in this work.

2.2. Direction of arrival estimation network (DOAnet)

Local shift-invariant features are extracted from the input spectrogram using CNN layers. In every CNN layer, the intra-channel time-frequency features are processed using a

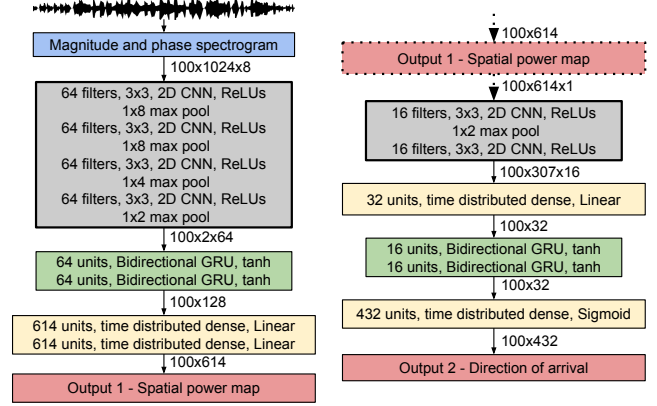


Fig. 1. DOAnet - neural network architecture for direction of arrival estimation of multiple sound sources.

receptive field of 3×3 , rectified linear unit (ReLU) activation and pad zeros to the resulting activation map to keep the output dimension equal to input. Batch normalization and max-pooling operation along frequency axis is performed after every CNN layer. The temporal information in the CNN activations is learned using RNN layers. Specifically, the bi-directional gated recurrent units (GRU) with tanh activation are used. The RNN output is mapped to the first output, the SPS, in regression manner using FC layers with linear activation.

The SPS is further mapped to DOA estimates—the final output of the proposed method—using a similar CRNN network as above with two minor architectural changes. An FC layer is introduced between the CNN and RNN layers to reduce the dimension of the RNN output. Additionally, the output layer which predicts the DOA uses sigmoid activation in order to estimate more than one DOA for a given time frame. Each node in this output layer represents a direction in 2-D polar space. During testing, the probabilities at these nodes are thresholded with a value of 0.5, so that anything greater suggests the presence of a source in the direction or otherwise absence of source.

We refer to the combined architecture of SPS and DOA estimation in this work as DOAnet. The DOAnet is trained using target SPS computed at each sampled direction, and for every time frame applying MUSIC (see Section 3.3), and is represented using nonnegative real numbers. For the DOA output, the DOAnet aims to make a discrete decision about the presence of a source in a certain direction; and during training, the DOAnet uses the ground truth DOAs utilized to synthesize the audio (see Section 3.1).

The DOAnet was trained for 1000 epochs using Adam optimizer, mean squared error loss for SPS output and binary cross entropy loss for DOA output. The sum of the two losses was used for back propagation. Dropout was used after every layer and early stopping was used if the DOA metric (Section 3.2) did not improve for 100 epochs. The DOAnet was implemented using Keras framework with Theano backend.

3. EVALUATION

3.1. Dataset

DOAnet is evaluated with datasets consisting of static point sources associated with a spatial coordinate in the space. We synthesize these datasets in two contexts - anechoic and reverberant; for each context, three datasets are generated with no temporally overlapping sources (*O1*), maximum two overlapping sources (*O2*), and maximum three overlapping sound sources (*O3*). We refer to the anechoic context dataset as *OxA* and reverberant as *OxR*, where x denotes the number of overlapping sources. Each of these datasets has three cross-validation (CV) splits with 240 recordings for training and 60 for testing. Recordings are sampled at 44.1 kHz and 30 s long.

In order to generate these datasets, we use the isolated real-life sound event recordings from the DCASE 2016 task 2 [20]. This dataset consists of 11 sound event classes, each with 20 examples. The classes in this dataset included speech, coughing, door slam, page-turning, phone ringing and keyboard sounds. During CV, for each of the splits, we randomly chose disjoint sets of 16 and 4 examples for training and testing, amounting to 176 examples for training and 44 for testing. In order to synthesize a recording, a random subset of the 176 or 44 sound examples was chosen from the respective split. The subset size varied for each recording based on the chosen sound examples. We start synthesizing a recording by randomly choosing the beginning time of the first sound example within the first second of the recording. The next sound example is placed 250-500 ms after the end of the first sound example. On reaching the maximum recording length of 30 s, the process is repeated as many times as the number of required overlapping sound events.

Each of the sound examples were assigned a DOA randomly using the following conditions. All sound events were placed in a spatial grid of ten degrees resolution along both azimuth and elevation. Two temporally overlapping sound events have at least ten degrees of spatial separation to avoid spatial overlapping. The elevation was constrained within the range of $[-60, 60]$ degrees, as most natural sound events occur in this range. Finally, for the anechoic dataset, the sound sources were randomly placed at a distance d in the range 1-10 m. For the reverberant dataset, the sound events were randomly placed inside a room of dimensions $10 \times 8 \times 4$ m with the microphone in the center of the room.

Spatialization for the anechoic case was done as following. Each point source signal s_i with DOA (ϕ_i, λ_i) , was converted to Ambisonics format by multiplying the signal with the vector $\mathbf{y}(\phi_i, \lambda_i) = [Y_{00}(\phi_i, \lambda_i), Y_{1(-1)}(\phi_i, \lambda_i), Y_{10}(\phi_i, \lambda_i), Y_{11}(\phi_i, \lambda_i)]^T$ of real orthonormalized spherical harmonics $Y_{nm}(\phi, \lambda)$. The complete anechoic sound scene multichannel recording \mathbf{x}_A was generated as $\mathbf{x}_A = \sum_i g_i s_i \mathbf{y}(\phi_i, \lambda_i)$, with the gains $g_i < 1$ modeling the distance attenuation. Each entry of \mathbf{x}_A corresponds to one channel and $g_i = \sqrt{1/10^{(d-1)/d_{max}}}$, where d_{max} is the maximum of $d = 10$.

In the reverberant case, a fast geometrical acoustics sim-

ulator was used to model natural reverberation based on the rectangular room image-source model [21]. For each point source s_i with DOA in the dataset, a sequence of K image sources was generated modeling reflections up to a predefined time-limit. Based on the room and its propagation properties, each image source was associated with a propagation filter h_{ik} and DOA (ϕ_k, λ_k) resulting in the spatial impulse response $\mathbf{h}_i = \sum_{k=1}^K h_{ik} \mathbf{y}(\phi_k, \lambda_k)$. The reverberant scene signal was finally generated by $\mathbf{x}_R = \sum_i s_i * \mathbf{h}_i$, where $(*)$ denotes convolution of the source signal with the spatial impulse responses. The room absorption properties were adjusted to match reverberation times of typical office spaces. Three sets of testing data were generated with similar room size as training data (Room 1), 80% of room size ($8 \times 8 \times 4$ m) and reverberation time (Room 2), and 60% of room size ($8 \times 6 \times 4$ m) and reverberation time (Room 3).

3.2. Metric

The DOAnet estimated SPS ($S_E(\phi, \lambda)$) is evaluated with respect to the baseline MUSIC estimated ground truth ($S_{GT}(\phi, \lambda)$) using the SNR metric calculated as $SNR = 10 \log_{10}(\sum_{\phi} \sum_{\lambda} S_{GT}(\phi, \lambda)^2 / \sum_{\phi} \sum_{\lambda} (S_E(\phi, \lambda) - S_{GT}(\phi, \lambda))^2)$.

As the DOA metric we use the angle between the estimate DOA (defined by azimuth ϕ_E and elevation λ_E) and the ground truth DOA $(\phi_{GT}, \lambda_{GT})$ in radians. This is calculated as $\sigma = \arccos(\sin \phi_E \sin \phi_{GT} + \cos \phi_E \cos \phi_{GT} \cos(\lambda_{GT} - \lambda_E))$. Further, to accommodate the scenario of unequal estimated and ground truth DOAs we calculate and report the minimum distance between them using the Hungarian algorithm [22] along with the percentage of frames in which the number of DOAs estimated were correct. The final DOA metric for the entire dataset is calculated by normalizing the minimum distance with the total number of estimated DOA's.

3.3. Baseline

The MUSIC pseudo-spectrum (S_{GT}) is based on a measure of orthogonality between the signal subspace (dominated by the source signals) of the spatial covariance matrix \mathbf{C} and the noise subspace (dominated by diffuse and ambient sounds, late reverberation, and microphone noise). The spatial covariance matrix is calculated as $\mathbf{C} = \mathbb{E}_{f,t} [\mathbf{X}(f, t) \mathbf{X}(f, t)^H]$, where \mathbf{X} is the multichannel spectrum, f, t are the frequency and time indices, H is the conjugate transpose and $\mathbb{E}_{f,t}$ denotes the expectation over f and t . For a sound scene with O number of sources, the S_{GT} is obtained from \mathbf{C} by first performing an eigenvalue decomposition on $\mathbf{C} = \mathbf{E} \mathbf{\Lambda} \mathbf{E}^H$. The sorted eigenvectors \mathbf{E} (according to eigenvalues with decreasing magnitude) are further partitioned into the two aforementioned subspaces $\mathbf{E} = [\mathbf{U}_s \ \mathbf{U}_n]$, where \mathbf{U}_s will be composed of O eigenvectors and the rest will form \mathbf{U}_n . The S_{GT} along the direction (ϕ_i, λ_i) is now given by $S_{GT}(\phi_i, \lambda_i) = 1/(\mathbf{y}^H(\phi_i, \lambda_i) \mathbf{U}_n \mathbf{U}_n^H \mathbf{y}(\phi_i, \lambda_i))$. Finally, the source DOAs are found by selecting the directions (ϕ_i, λ_i) corresponding to the O largest peaks from S_{GT} .

Table 1. Evaluation metric scores for the spatial power map and DOAs estimated by the DOAnet for different datasets.

	Anechoic			Reverberant (Room 1)		
Overlap upto	1	2	3	1	2	3
SPS (in dB)	9.90	3.35	-0.26	3.11	1.24	0.13
DOA error with unknown number of active sources (threshold of 0.5)						
DOAnet	0.01	0.14	0.32	0.11	0.20	0.67
Correctly predicted frames (in %)	0.95	0.42	0.01	0.59	0.15	0.01
DOA error with known number of active sources						
DOAnet	0.02	0.48	0.86	0.22	0.68	1.17
MUSIC	0.04	0.15	0.50	0.45	1.00	1.60

3.4. Evaluation procedure

The parameter tuning for DOAnet was performed on the *O1A* test data, and the best configuration is as shown in Figure 1. This configuration has 677 K weights, and the same configuration is used in all of the following studies.

At test time, the SNR metric for SPS output of the DOAnet (S_E) is calculated with respect to SPS of baseline MUSIC (S_{GT}). The DOA metric for the DOAs predicted by DOAnet and baseline MUSIC are calculated with respect to the ground truth DOA used to synthesize the dataset.

In the above experiment, the baseline MUSIC algorithm uses the knowledge of the number of active sources. In order to have a fair evaluation, we test the DOAnet in a similar scenario where the number of sources is known. We use this knowledge to choose the top probabilities in prediction layer of the DOAnet instead of thresholding it with a value of 0.5.

4. RESULTS AND DISCUSSION

The results of the evaluations are presented in Table 1. The high SNRs for SPS in both the contexts, with up to one and two overlapping sound events show that the SPS generated by DOAnet (S_E) is comparable with the baseline MUSIC SPS (S_{GT}). Figure 2 shows the S_E and the respective S_{GT} when two active sources are closely located. In the case of up to three overlapping sound events, the baseline MUSIC is already at its theoretical limit of estimating $N - 1$ sources from N -dimensional signal space [23]. In practice, for $N - 1$ sources only one noise subspace vector U_n is used to generate SPS, which for real signals is too weak for stable estimation. With more than four channels input, in other words, with higher order Ambisonic signals we should be able to localize more sound sources.

The DOA error for the proposed DOAnet when the num-

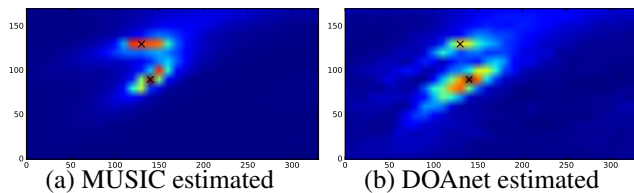


Fig. 2. SPS for two closely located sound sources. The black-cross markers represent the ground truth DOA. The horizontal axis is azimuth and vertical axis is elevation angle (in degrees)

Table 2. Evaluation scores for unmatched reverberant room.

	Room 2		Room 3	
Overlap upto	1	2	1	2
SPS (in dB)	3.53	1.49	3.49	1.46
DOAnet error (Unknown number of sources)				
DOAnet	0.06	0.12	0.08	0.19
Correctly predicted frames (in %)	0.46	0.14	0.49	0.14
DOAnet error (Known number of sources)				
DOAnet	0.15	0.56	0.16	0.59
MUSIC	0.55	1.02	0.58	1.06

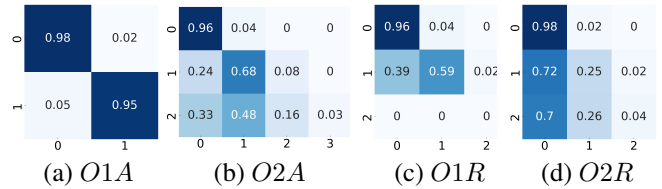


Fig. 3. Confusion matrix for the number of DOA estimated per frame by the DOAnet. The horizontal axis is the DOAnet estimate, and the vertical axis is the ground truth.

ber of active sources are unknown is presented in Table 1. The DOAnet error is considerably better in comparison to the baseline MUSIC that uses the active sources knowledge for all datasets. However, the number of frames in which DOAnet produced exact number of active sources were few. For example, in the case of anechoic recordings with up to two overlapping sound events, only 42% of the estimated frames had the correct number of DOA predictions. This prediction drops even drastically when the number of sources is three, due to the theoretical limit of MUSIC as explained previously, and consequently for the DOAnet as MUSIC SPS is used for training. The confusion matrix for the number of DOA estimates per frame for different datasets is visualized in Figure 3.

With the knowledge of the number of active sources (Table 1), the DOAnet performs considerably better than baseline MUSIC for all datasets other than the *O2A* and *O3A*. This shows that the proposed DOAnet has the potential to learn the directional information of sound sources directly from the spectrogram of the input audio. Further, the DOAnet error on unmatched reverberant data is presented in Table 2. The performance of DOAnet is seen to be consistent in comparison to the matched reverberant data in Table 1, and significantly better than the performance of MUSIC.

5. CONCLUSION

A convolutional recurrent neural network (DOAnet) was proposed for multiple source localization. The DOAnet was shown to learn the number of active sources directly from the input spectrogram, and estimate precise DOA in 2-D polar space. The method was evaluated on anechoic, matched and unmatched reverberant dataset. The proposed DOAnet performed considerably better than baseline MUSIC in most scenarios. Thereby showing the potential of DOAnet in learning highly computational algorithm without prior knowledge of the number of sources.

6. REFERENCES

- [1] Matthias Woelfel and John McDonough, "Distant speech recognition," in *Wiley*, 2009.
- [2] Tara Sainath et al, "Multichannel signal processing with deep neural networks for automatic speech recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2017.
- [3] Archontis Politis, Juha Vilkkamo, and Ville Pulkki, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866, 2015.
- [4] Joonas Nikunen and Tuomas Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.
- [5] Yiteng Huang, J. Benesty, G.W. Elko, and R.M. Mersereau, "Real-time passive source localization: a practical linear-correction least-squares approach," in *IEEE Transactions on Speech and Audio Processing*, 2001.
- [6] Michael S. Brandstein and Harvey F. Silverman, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997.
- [7] Ralph O Schmidt, "Multiple emitter location and signal parameter estimation," in *IEEE Transactions on Antennas and Propagation*, 1986.
- [8] Richard Roy and Thomas Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," in *IEEE Transactions on Audio, Speech, and Language Processing*, 1989.
- [9] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180. 2001.
- [10] Reinhild Roden, Niko Moritz, Stephan Gerlach, Stefan Weinzierl, and Stefan Goetze, "On sound source localization of speech signals using deep neural networks," in *Deutsche Jahrestagung für Akustik (DAGA)*, 2015.
- [11] Xiong Xiao et al, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [12] Ryu Takeda and Kazunori Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016.
- [13] Alfredo Zermine, Wenwu Wang, Qiuqiang Kong, Yong Xu, and Mark D. Plumbley, "Deep neural network based audio source separation," in *International Conference on Mathematics in Signal Processing*, 2016.
- [14] Fabio Vesperini et al, "A neural network based algorithm for speaker localization in a multi-room environment," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.
- [15] Soumitro Chakrabarty and Emanuel A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.
- [16] Heinz Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, vol. 348, Springer, 2007.
- [17] Tara N. Sainath, Oriol Vinyals, Andrew Senior, and Hasim Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [18] Miroslav Malik et al, "Stacked convolutional and recurrent neural networks for music emotion recognition," in *Sound and Music Computing Conference (SMC)*, 2017.
- [19] Sharath Adavanne, Pasi Pertilä, and Tuomas Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
- [20] Emmanouil Benetos, Mathieu Lagrange, and Gregoire Lafay, "Sound event detection in synthetic audio," <http://www.cs.tut.fi/sgn/arg/dcass2016/>, 2016.
- [21] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small-room acoustics," in *The Journal of the Acoustical Society of America*, 1979.
- [22] Harold W. Kuhn, "The hungarian method for the assignment problem," in *Naval Research Logistics Quarterly*, 1955, number 2, p. 8397.
- [23] Bjorn Ottersten et al, "Exact and large sample maximum likelihood techniques for parameter estimation and detection in array processing," in *Radar Array Processing. Springer Series in Information Sciences*, 1993.