PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Source localization in reverberant rooms using Deep Learning and microphone arrays

Hadrien PUJOL*, Éric BAVU*, and Alexandre GARCIA*

*Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC), Conservatoire national des arts et métiers (Cnam), 292 rue Saint-Martin 75003 Paris, France

Abstract

Sound sources localization (SSL) is a subject of active research in the field of multi-channel signal processing since many years, and could benefit from the emergence of data-driven approaches. In the present paper, we present our recent developments on the use of a deep neural network, fed with raw multichannel audio in order to achieve sound source localization in reverberating and noisy environments. This paradigm allows to avoid the simplifying assumptions that most traditional localization methods incorporate using source models and propagating models. However, for an efficient training process, supervised machine learning algorithms rely on large-sized and precisely labelled datasets. There is therefore a critical need to generate a large number of audio data recorded by microphone arrays in various environments. When the dataset is built either with numerical simulations or with experimental 3D soundfield synthesis, the physical validity is also critical. We therefore present an efficient tensor GPU-based computation of synthetic room impulse responses using fractional delays for image source models, and analyze the localization performances of the proposed neural network fed with this dataset, which allows a significant improvement in terms of SSL accuracy over the traditional MUSIC and SRP-PHAT methods.

Keywords : Localization, Deep-Learning, Dataset, Reverberant

1 INTRODUCTION

Sound sources localization (SSL) in reverberant or noisy environments is a challenging task, with wide applications ranges, such as sensitive areas acoustic surveillance, speaker localization for videoconferencing tasks, acoustic diagnosis in industrial engineering, or acoustic survey systems of animal species. In order to solve this inverse problem using acoustic measurements, many classes of algorithms have been developed in the last decades using microphone array signal processing, propagation models, and signal models. However, when the background noise is high and the propagating medium differs from the model, these acoustic localization methods are either less efficient or require significant computation time.

In the last years, Deep Learning methods have been demonstrated to be efficient for a large number of tasks, especially for computer vision and machine hearing. More specifically, for acoustic based applications, Deep Learning approaches have been successfully applied for automatic speech recognition (ASR) or speaker recognition and allow to achieve state of the art performances. More recently, the use of Deep Learning methods and data-based paradigms have been explored by the scientific community for sound sources localization tasks. In order to reach this goal, several promising approaches have been proposed. Some are them are based on pre-processed features calculated from the measured multichannel signals, either in the frequency domain [1]–[4] or in the ambisonic domain [5]. In this paper, we propose to perform a SSL task using raw multichannel audio signals. The main intent is to perform the localization inference in real-time, without having to rely on hand-crafted features or signal based hypothesis. We therefore propose a neural network architecture that aims at performing a joint feature learning process, in order to solve the SSL inverse problem from raw multichannel audio, in a similar way to what we recently proposed for sound recognition tasks using raw audio [6].

For data-based methods, the learning efficiency strongly relies on the dataset used for the training process. For a SSL task, several approaches can be exploited in order to build the multichannel audio dataset. These ap-

proaches can be either based on realistic numerical simulations of the direct problem, on extensive experimental measurements, or on hybrid methods which use real-life (or simulated) room impulse responses (RIRs) in order to perform a convolution with real-life recordings [7]. For simulation-based or hybrid-based approaches, the dataset has to be realistic enough in order to match real recording conditions. In addition, a large variability in the dataset should be available during the learning process, so that features can be efficiently extracted from the raw multichannel data without overfitting to a specific configuration. We therefore developed an efficient tensor GPU-based computation of synthetic RIRs, using fractional delays for image source models. Using this dataset for a compact microphone array, the proposed neural network architecture allows to perform a 2D DOA inference with a significant improvement in terms of SSL accuracy of 1^o to 25^o in anechoic and reverberant environments and strong noisy conditions over the MUSIC and SRP-PHAT methods, for speech signals.

2 METHODS AND EVALUATION

2.1 Multichannel reverberant audio Dataset

In the context of the development of this Deep Learning based approach for SSL using raw audio data, we have built several datasets using the MiniDSP UMA-8 compact microphone array (see https://www.minidsp.com/products/usb-audio-interface/uma-8-microphone-array) geometry, either using simulated data or measured data. This geometry matches those of the microphone arrays used in personal assistants such as Amazon Alexa, Google Home, or Apple HomePod. The UMA-8 array is composed of 7 digital MEMS microphones, the first one being at the center of the array, and the 6 others being evenly distributed on a 8 cm diameter circle. For both the dataset development and the neural network architecture, the proposed approach has been tailored in order to acomodate any microphone array configurations. However, in this paper, for sake of clarity and concision, we chose to illustrate the results using this particular array, both for model-based and data-based approaches.

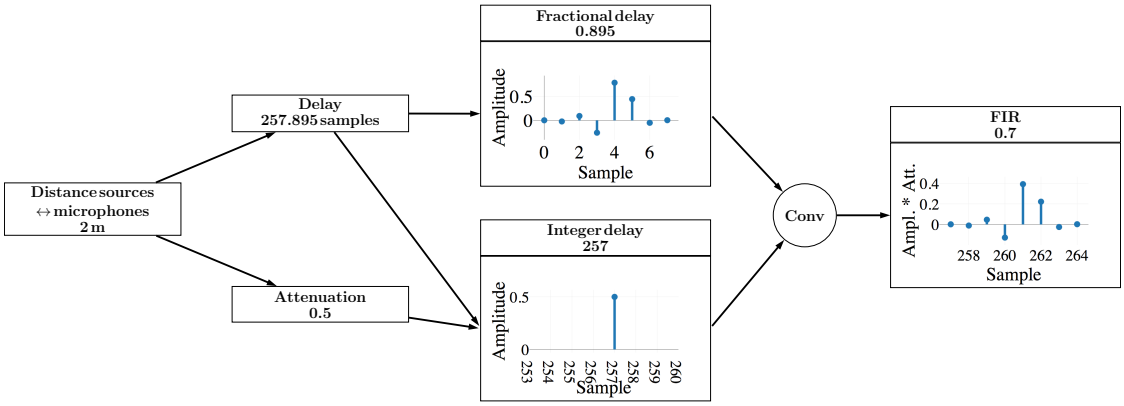
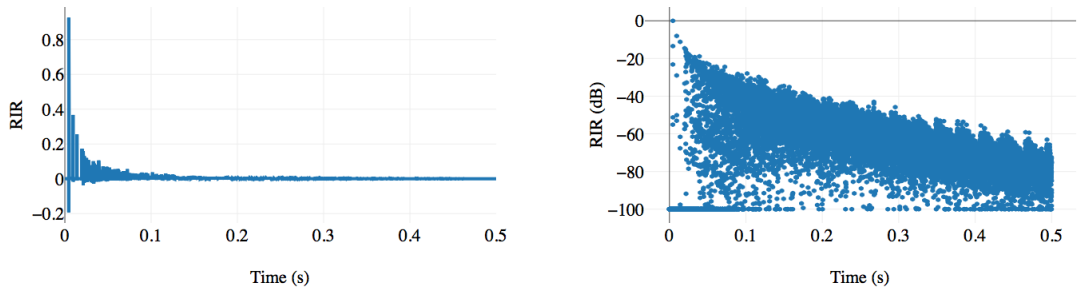


Figure 1. 7^{th} order Lagrange interpolation for the fractional delay computation, for a single source/microphone. This approach is used for more than 20×10^9 image source-receiver configurations in order to compute the 48000×7 RIRs used to build the dataset corresponding to a typical classroom geometry, with a T_R of 0.5 s.

The multichannel reverberant dataset relies on the computation of realistic RIRs using the image method [8], for 48000 source positions in the considered room, and for every microphone position of the compact measurement array. As a consequence, the dataset relies on the computation of 336000 RIRs for this particular microphone array. For each of these RIRs, the whole number of image sources positions and corresponding attenuations that contribute to the acoustic field for a time interval of $[0; T_R]$ are computed using the Pyroomacoustics library [9]. For a typical classroom size of 7x10x3.70 m with a T_R of 0.5 s, the whole number of image sources therefore represents more than 80000 image sources for each RIR.

For such a large number of RIRs and image sources, the existing frameworks did not allow to perform the RIR computation efficiently and accurately enough. This is the reason why we developed for this specific application a fast parallel batch RIR computation performed on the GPU using the Tensorflow APIs [10]. This implementation is achieved using sparse tensors computations and an efficient fractional delay filters implementation (see Fig. 1). For a compact microphone such as the UMA-8, there is a critical need to keep the precision of the time delays corresponding to the distance between each of the image sources involved in the RIR computation and the microphones. In order to implement the RIR with non integer sample delays, we therefore used a 7^{th} order Lagrange interpolation [11], that allows to perform a better accuracy that standard truncated sinc interpolations, with a much lower computation cost. The integer part of the sample delay is represented as a sparse array, and the 7 coefficients of each fractional delay filters along with the individual dampings corresponding to the cumulative effect of each wall of the room and the distance between the image source and the microphones are then used (see Fig. 1) to compute the RIRs using every source image contributions (see Fig. 2). Using this large number of RIRs, the complete dataset is built using batch convolutions with real-life recordings.



(a) Linear representation

(b) Logarithmic representation

Figure 2. Example of the simulated RIR for a source in a room with a reverberation time of 0.5 s using the proposed GPU-based fast implementation.

2.2 Proposed neural network architecture

The proposed neural network architecture is directly fed with raw multichannel audio data measured by the microphone array (see Fig. 3), without any further preprocessing, in order to allow a joint feature learning along with the SSL task. We also specifically developed this architecture in order to allow real-time SSL inference after convergence of the training process. The convolutional neural cells, which are widely used in deep neural network architectures, can be thought as a strict equivalent to learnable finite impulse response (FIR) filtering in standard digital signal processing (DSP). Conventional beamforming methods can be thought as a filter and sum approach [12]. The global neural architecture depicted on Fig. 3 has been therefore developed in a similar way, with specific operations allowing the neural network to be expressive enough to achieve the SSL task in reverberant and noisy environments using raw multichannel audio data.

The proposed neural network is built using successive subnetworks that act as learnable filterbanks. Each subnetwork is implemented as a residual network of 1D depthwise separable atrous convolutional layers with exponentially increasing dilation factors, which have recently emerged as an efficient architecture for audio generation [13], denoising [14], neural translation [15], and raw audio recognition [6]. The atrous convolutions are performed independently over every input channel, and allow to achieve a large receptive field with only 6 sets of one-dimensional convolutions with kernels of size 1×3. This presents the considerable advantage of making a much more efficient use of the parameters available for representation learning than standard convolutions [15].

These computed depthwise convolutions are then projected onto a new channel space for each layer using a pointwise convolution. Each atrous convolution layer is followed by a layer normalization process and a tanh activation function. After M successive filterbanks ($M = 3$ on Fig. 3), a pseudo-energy computation is achieved using a mean computation of the squared output channels followed by a Selu activation function [16]. This nonlinear activation function has been recently introduced in the literature in order to avoid standard batch normalization processes, without degrading the computational efficiency of deep neural networks. The computed pseudo-energies are then fed to a dense layer in order to compute the overall output of the neural network, which corresponds to the estimation of the source's position in cartesian coordinates, similarly to what Adavanne et al. proposed in [1], in order to avoid any ambiguity due to discontinuies at the angular wrap-around for DOA retrieval, which is treated as a regression task.

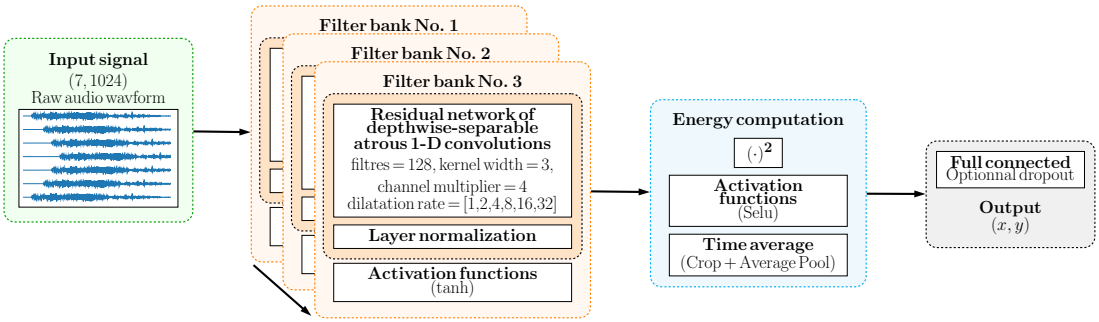


Figure 3. Global neural network architecture for the SSL task, based on successive learnable filterbanks based on residual subnetworks of dephwise separable atrous 1D convolutions with increasing dilation rates and pseudo-energy computation. The output corresponds to the estimated source position in cartesian coordinates.

The cost function is computed using the mean square error loss on the spherical distance between the reference and the estimated locations, in order to solve the SSL as a regression task. In order to enforce overall energy conservation between the measured input channels and the output of the successive learnable filterbanks, a second term is added to this cost function in order to minimize the L1-norm of the difference between these energies. The learning and backpropagation of errors through the neural network is optimized using the Adaptive Moment Estimation (Adam) [17] algorithm, which performs an exponential moving average of the gradient and the squared gradient, and allows to control the decay rates of these moving averages. The models have been implemented and tested using the Tensorflow open source software library [10].

2.3 Methods

As explained in the previous section, the dataset is computed using a large number of RIRs, which are batch convoluted with real life recordings, similarly to what has been proposed in [4], [5], [7]. The dataset is split into training, validation and test sets, with a ratio of 80:10:10. The proposed approach has been tested both in an anechoic environment and in a reverberant environment with a typical classroom size of $7 \times 10 \times 3.7$ m and a reverberation time of 0.5 s. In this case, the microphone array is positioned at coordinates $[4, 6, 1.5]$ m, the origin of the Cartesian coordinate system corresponding to a corner of the room.

For each environments, the source positions are randomly drawn from a uniform distribution in a torus of mean radius $R = 2$ m, centered on the microphone array position, with a section dR.Rsin($d\phi$) (see Fig. 4). Since we aim at achieving 2D-DOA retrieval of the angular position θ, this process allows to introduce a spatial variability in order to increase the robustness of the proposed method to unseen conditions.

For each environment, the dataset is equally distributed over 6 kinds of real life signals convoluted with 8000 random positions, each signal corresponding to different kinds of frequency contents : a 1000 Hz sinusoidal tone, a car horn, anechoic recordings (see https://users.aalto.fi/~ktlokki/Sinfrec/sinfrec.html) of

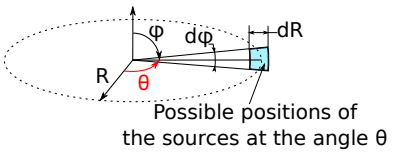


Figure 4. Random distribution of the source positions at a given angular position θ

symphonic music [18], and outdoor recordings of a people's multiple talking in Danish (mostly women) (see https://odeon.dk/downloads/anechoic-recordings/). At the training phase, each multichannel recordings of the training mini batches are added to a random Gaussian noise with randomly picked signal-to-noise ratio (SNR) values in the range of $[+\infty, +10]$ dB. This process can be interpreted as a data augmentation strategy, which allows to ease the generalization capabilities of the proposed method during the learning process.

2.4 Evaluation

Since the output of the proposed neural network corresponds to continuous cartesian coordinates (x, y), the estimated angular position of the sources are inferred using $\theta = \frac{180}{\pi} \arctan 2(\frac{y}{x})$. In order to evaluate the 2D-DOA SSL performances of the proposed algorithm and to compare it in a reproducible way to model-based algorithms (MUSIC [19] and SRP-PHAT [20]), each method is tested using the exact same testing dataset. This testing dataset consists in 360 sources, uniformly distributed over a circle of radius $R = 2$ m, using speech data (unseen during the training phase). The mean DOA mismatch is evaluated over these 360 positions, for each proposed environments, and for different deterministic (SNR) values, ranging from $+\infty$ (no noise) to 0dB in order to evaluate the robustness of each methods to noisy measurements.

3 RESULTS AND DISCUSSION

In this section, the SSL performances of the proposed method, denoted as "BeamLearning" are compared with those obtained using the MUSIC and SRP-PHAT methods. The performance metric used here corresponds to the mean absolute angular mismatch over the whole 360 test positions.

3.1 Ideal case of non-noisy measurements : robustness to reverberation

In order to assess the capabilities of the BeamLearning approach, the proposed deep neural network has been trained using two different datasets, corresponding to an anechoic environment, and the reverberant environment with a reverberation time of 0.5 s described in the previous sections. The neural network has been trained using the proposed data augmentation, with a random SNR value for each element of the training mini-batches. Even if the neural network already allows to obtain good results after a small number of epochs (50 epochs for the anechoic case and 250 epochs for the reverberant case), the training phase has been performed during 2600 epochs, which improved SSL performances by a few tenths of degrees, without any sign of overfitting.

Method	Anechoïc room	Reverberant room ($T_R = 0.5$ s)
MUSIC	**0.25°**	3.6°
SRP-PHAT	**0.25°**	2.5°
BeamLearning (proposed)	0.35°	**2.4°**

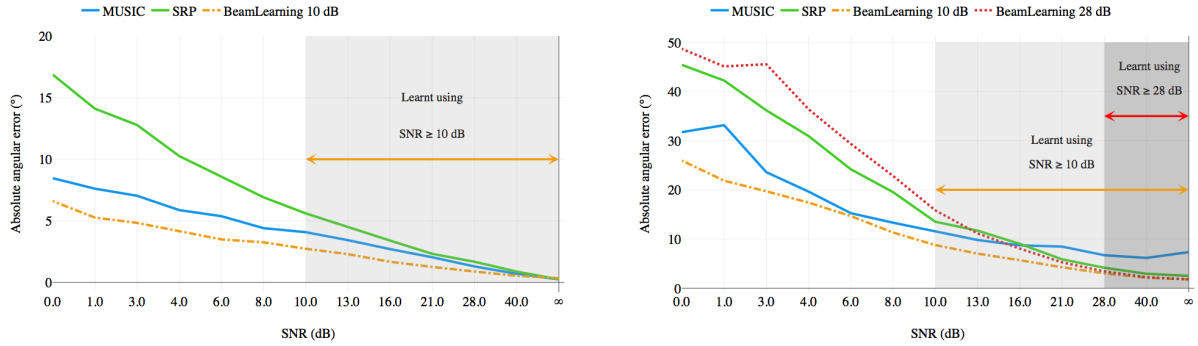
Table 1. Mean absolute angular mismatch obtained using the speech testing dataset (360 positions) for the two model-based methods and the proposed data-based methods for a SSL task in the ideal case of non-noisy measurements. The proposed neural network has been trained using noisy measurements, with random SNR values ranging from $+\infty$ to $+10$ dB. Bold font indicates best result obtained.

Table 1 shows the obtained results using MUSIC, SRP-PHAT, and BeamLearning, after convergence of the learning process. Interestingly, these results show that the BeamLearning approach performs significantly better than both model-based methods in the reverberant environment. In the simpler case of an anechoïc room however, the obtained angular mismatch obtained using the proposed approach is slightly higher than with MUSIC and SRP-PHAT, but the DOA estimation error remains very low for the three methods.

Interestingly, when the neural network is trained using non-noisy data (no data augmentation), the BeamLearning approach allows to obtain a DOA mismatch of 0.28^{o} in an anechoïc room, which matches the performances of MUSIC and SRP-PHAT. As shown in the following section, a further exploration of the SSL performances in noisy environments shows that the data augmentation method allows BeamLearning to be more robust to noise than the model-based methods, which explains the reasons why Table 1 shows the testing results obtained using data augmentation, even if the obtained results are slightly worse for the ideal case of non-noisy measurements.

3.2 Robustness to measurement noise

These competitive results obtained for the SSL task in ideal conditions using the BeamLearning approach motivated a systematic analysis of performances of the proposed deep neural network after convergence under noisy conditions, including heavy cases, where the SNR value equals to 0 dB (the data augmentation during the training phase only included higher SNR values than +10 dB). The obtained results are presented on Fig. 5, both for the anechoïc case and the reverberant case.



(a) Anechoïc environment

(b) Reverberant environement ($T_R = 0.5$ s)

Figure 5. Mean absolute angular mismatch obtained using the speech testing dataset (360 positions) with SNR values ranging from $+\infty$ to 0 dB. These DOA errors are plotted for the MUSIC method (solid blue curve), the SRP-PHAT method (solid green curve), and the proposed BeamLearning approach (dotted curves), in both an anechoïc environment (a) and a reverberant environment (b). The proposed neural network has been trained using noisy measurements, with random SNR values ranging from $+\infty$ to $+10$ dB (orange dash-dotted curve) and random SNR values ranging from $+\infty$ to $+28$ dB (red dotted curve).

In both environments, the obtained results show that the BeamLearning approach allows to achieve better SSL performances under noisy conditions than the model-based algorithms, even for lower SNR values than those used in the training phase. As expected, SRP-PHAT gives the worst performances for noisy measurements, and MUSIC allows to improve the DOA estimation. However, the BeamLearning approach allows to improve the DOA estimation precision by 1^{o} to 10^{o} over the MUSIC method (resp. 10^{o} to 24^{o} over the SRP-PHAT method), for SNR values lower than 20 dB in reverberant environments. In the anechoic case, the same trend is observed, with a 2^{o} to 4^{o} improvement over the MUSIC method, and a 3^{o} to 10^{o} improvement over the SRP-PHAT method, for SNR values lower than 20 dB.

In order to further investigate the influence of the data augmentation process, the neural network has also been

trained in reverberant environments, with added random Gaussian noise with randomly picked SNR values in the range of $[+\infty, +28]$ dB instead of $[+\infty, +10]$ dB (see dotted red curve on Fig. 5). The obtained results show that in this case of less intensive data augmentation, the BeamLearning approach is less robust to noise for SNR values lower than 15 dB, therefore validating the proposed approach of data augmentation with a high variability in SNR values, in order to increase robustness to measurement noise for SSL assessment.

3.3 Computational efficiency

As shown in the previous sections, the proposed Deep Learning approach strongly relies on a large sized dataset, with a high variability in terms of sources positions, signals, and SNR values. Using this process, the BeamLearning approach allows to achieve the SSL task with a higher precision and a better robustness to measurement noise than MUSIC and SRP-PHAT methods, in both reverberant and anechoïc environments, including under unseen heavy noise conditions. Using the proposed datasets, the whole learning process took approximately 190 hours of computation using a Nvidia GTX 1080Ti GPU card, for a total of 645 hours of audio waveforms processed by the proposed model. This computation time in the learning phase includes the feed forward propagation, loss function computations, back-propagation, gradients computations and variables updates using Adam. For inference at the testing phase however, the neural network model parameters are frozen, and the only computations involved correspond to the feed forward propagation of the input data to evaluate the source's angular position. In this case, the mean computation time for one DOA estimation reduces to 0.2 ms on the same GPU card, which is significantly faster than the mean computation times of 660 ms and 47 ms for the MUSIC and SRP-PHAT methods using an Intel Core i7-6900K CPU card.

4 CONCLUSIONS

In this paper, we presented BeamLearning, a machine learning approach for the sound source localization task, using raw multichannel audio. The proposed neural network allows to estimate a sound source's angular position, with a higher accuracy and a better robustness to noisy measurements than the traditional model-based MUSIC and SRP-PHAT methods. The proposed data-based approach is inspired by the filter and sum approach used in traditional beamforming, and allows to estimate a source position in real time, using a neural network which mainly consists in successive learnable filterbanks based on residual subnetworks of depthwise separable atrous 1D-convolutions, followed by a pseudo-energy computation over the learnt channels. We also show that the success of this approach strongly relies on the dataset used in the training phase, which has been built using an efficient implementation on the GPU of a large number of RIR, and a batch convolution on the GPU of these computed RIRs with real life recordings. In a future work, an experimental dataset will be presented, based on computed RIRs and the use 3D sound field synthesis using higher ambisonics. Preliminary results on this experimental dataset show that similar improvements are observed, with the main advantage of allowing the neural network to build the best representation in order to compensate the individual microphone frequency responses and the diffraction effects induced by the microphone array structure, which are not taken into account using standard model-based methods.

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 Journ. of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[2] S. Chakrabarty and E. A. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals", in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*, IEEE, 2017, pp. 136–140.

[3] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals", *IEEE Journ. of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[4] J. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: from audio signals to source position coordinates", *Sensors*, vol. 18, no. 10, p. 3418, 2018.

[5] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings", *IEEE Journ. of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[6] É. Bavu, A. Ramamonjy, H. Pujol, and A. Garcia, "TimeScaleNet : a Multiresolution Approach for Raw Audio Recognition using Learnable Biquadratic IIR Filters and Residual Networks of Depthwise-Separable One-Dimensional Atrous Convolutions", *IEEE Journ. of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 220–235, 2019.

[7] S. Adavanne, A. Politis, and T. Virtanen, "A Multi-room Reverberant Dataset for Sound Event Localization and Detection", in *Submitted to Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019. [Online]. Available: https://arxiv.org/abs/1905.08546.

[8] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics", *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[9] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: a python package for audio room simulation and array processing algorithms", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 351–355.

[10] M. Abadi, A. Agarwal, P. Barham, *et al.*, *TensorFlow: large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: http://tensorflow.org/.

[11] J. O. Smith, *Introduction to Digital Filters with Audio Applications*. W3K Publishing, 2007, ISBN: 978-0-9745607-1-7.

[12] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[13] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: a generative model for raw audio", *arXiv preprint arXiv:1609.03499*, 2016.

[14] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5069–5073.

[15] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise separable convolutions for neural machine translation", in *International Conference on Learning Representations*, 2018, pp. 1–10.

[16] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks", in *Advances in Neural Information Processing Systems*, 2017, pp. 971–980.

[17] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization", in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[18] T. Lokki, J. Patynen, and V. Pulkki, "Recording of anechoic symphony music", *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3936–3936, 2008.

[19] R. Schmidt, "Multiple emitter location and signal parameter estimation", *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[20] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms", in *Microphone Arrays*, Springer, 2001, pp. 157–180.