INVESTIGATING THE GENERALIZATION ABILITIES OF A DEEP LEARNING METHOD FOR SOUND SOURCE LOCALIZATION USING SMALL-SIZED MICROPHONE ARRAYS

Giovanni Affatato, Roberto Alessandri

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20122 Milano, Italy
[giovanni.affatato,roberto.alessandri]@mail.polimi.it

ABSTRACT

In recent years, Deep Learning (DL) methods helped to overcome trade-offs constraining signal processing tasks. Sound Source Localization (SSL) in noisy and reverberant conditions, especially with small-sized microphone arrays, is one of them. Yet, the success of DL techniques depends vastly on data, which influences the ability of the model to abstract from the peculiarities of the training dataset. In this work, we test such abilities on a Convolutional Neural Network (CNN) trained specifically to infer the Direction Of Arrival (DOA) of a sound source in these conditions. To this end, we generated three datasets via simulation techniques that address different acoustic parameters: the volume of the room, the position of the microphone array in the room and the distance of the source. We show that the CNN is able to perform well on the first two under the condition of augmenting the dataset.

Index Terms— Deep Learning, Differential Microphone Arrays, Sound Source Localization, Sound Intensity, Convolutional Neural Networks

1. INTRODUCTION

Sound Source Localization is the problem of estimating the Direction Of Arrival of a source from multiple signals acquired by a microphone array. The importance of solving this problem is given by its numerous applications, ranging from hearing aids [1], teleconferencing [2], source separation [3], speech recognition [4] to many others.

Classical methods for SSL are rooted in the field of signal processing. Some of the most popular approaches are based on spatial filtering (DAS [5], MUSIC [6], ESPRIT [7]), on the Time Difference of Arrival (GCC [8]) and on the Steered Response Power [9][10]. However, it is known that these methods generally do not perform well in presence of noise and reverberation. In addition, it is important to take into consideration the aperture size of the microphone array. Since a larger aperture size corresponds to higher spatial resolution [11] and therefore better performances. But in some applications the choice of the aperture size is limited due to the small space in which the microphone array is placed (e.g. smartphones, headphones).

In recent years, following the success achieved in other fields such as computer vision, Deep Learning methods have been proposed to address the aforementioned issues. The main advantage is to incorporate in the training process information about the acoustic properties of the room. While classical methods are based only on the spatial information embedded in the configuration of the array. As a result, data-driven methods such as DL can outperform classical methods, at the cost of dealing with a large amount of data, real or simulated. The effectiveness of a DL model is measured by its ability to generalize over different dimensions of the application (e.g. noise levels, reverberation times and so on). Namely, to correctly classify new data with characteristics that differ from the ones learned during training.

In our work, we investigate the generalization capabilities of two CNN which are state-of-the-art for classifying signals recorded by a small-sized microphone array composed of two orthogonal first-order Differential Microphone Arrays (DMAs) [12]. The CNNs are structurally equal except for the fact that one is able to localize locations with a resolution of 30° and the other 10°. Thus, we reproduced the same feature extraction scheme based on Sound Intensity (SI) estimation and the CNN architecture are described in [12]. Then, we generated three datasets with simulation techniques to test how the models behave if we change the volume of the room, the position of the microphone array in the room and the distance of the sources. We demonstrate that the models are able to generalize over the first two parameters. While the experiments over the third parameter suggest that this method is not well suited.

Our report is structured as follows. Section 2 describes the signal model and the SI estimation. Section 3 describes the feature extraction scheme and the CNN architecture. Section 4 describes the details of the simulations and Section 5 the results. In Section 6 we draw our conclusions.

2. SIGNAL MODEL AND BACKGROUND

2.1. Microphones Configuration and Signal Model

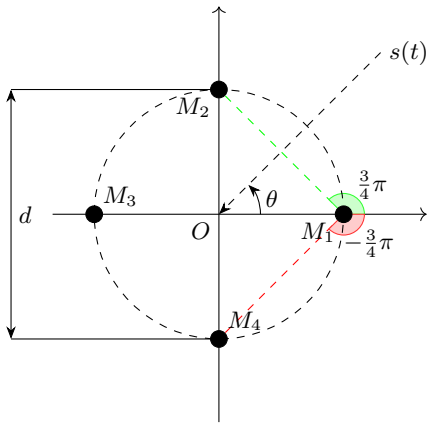


Figure 1: Configuration of the microphone array

The microphone array considered in [12] is composed by four omnidirectional microphones arranged as shown in Figure 1.

It can be decomposed into two orthogonal first-order DMAs of size $d = 0.04\ m$ with center in the origin O: $M_1 - M_3$ and $M_2 - M_4$. The DOA θ of a sound source is the angle defined with respect to the positive x-axis, thus $\theta \in [-180, 180)^\circ$. The signal acquired by a microphone can be expressed as

$$p_i(t) = h_i(t) * s(t) + n_i(t), \tag{1}$$

where $s(t)$, $h_i(t)$ and $n_i(t)$ are respectively the source signal, the Room Impulse Response (RIR) and the additive noise, while $*$ is the convolution.

2.2. Sound Intensity Estimation

The Sound Intensity is defined as the power carried by sound waves per unit area in a direction r perpendicular to that area, therefore is the quantity that contains the location information of our source. We start from the following equation that relates the particle velocity $v_r(t)$ in the direction r and the sound pressure $p(t)$ [13]

$$v_r(t) = -\frac{1}{\rho}\int_{-\infty}^{t} \frac{\partial p(\tau)}{\partial r} d\tau, \tag{2}$$

where ρ is the air density. The pressure gradient can be approximated by a finite difference [14] between two close points separated by a distance Δr along the direction r

$$v_r(t) \approx -\frac{1}{\Delta r \rho}\int_{-\infty}^{t} [p_{r_2}(t) - p_{r_1}(t)] d\tau. \tag{3}$$

The two pressures $p_{r_1}(t)$ and $p_{r_2}(t)$ in eqn. (3) can be considered as the two inputs of a first-order DMA. Hence, following the orientation of $M_1 - M_3$ and $M_2 - M_4$, we can express eqn. (3) in the Time-Frequency (T-F) domain as

$$V_x(\omega, t) = j\frac{[P_3(\omega, t) - P_1(\omega, t)]}{\omega \rho d}, \tag{4}$$

$$V_y(\omega, t) = j\frac{[P_4(\omega, t) - P_2(\omega, t)]}{\omega \rho d}, \tag{5}$$

where $P_i(\omega, t)$ is the Short-Time Fourier Transform (STFT) of the measured signal at microphone M_i and $j = \sqrt{-1}$ is the imaginary unit. Lastly, the sound pressure in the origin O can be estimated by averaging the received signals from all microphones [14], in the T-F domain:

$$P_0(\omega, t) = \frac{P_1(\omega, t) + P_2(\omega, t) + P_3(\omega, t) + P_4(\omega, t)}{4}. \tag{6}$$

According to [15], we can express the x and y components of the instantaneous complex SI as

$$I_{0x}(\omega, t) = P_0(\omega, t) V_x^\star(\omega, t), \tag{7}$$

$$I_{0y}(\omega, t) = P_0(\omega, t) V_y^\star(\omega, t), \tag{8}$$

where * indicates the complex conjugate.

3. FEATURE EXTRACTION AND NETWORK ARCHITECTURE

In this section, we recap for convenience the feature extraction process and the network architecture proposed in [12].

3.1. Feature Extraction

The location information is embedded in the real part of the complex SI [15], thus the extracted SI features can be expressed as

$$I_x(\omega, t) = \Re[I_{0x}(\omega, t)], \tag{9}$$

$$I_y(\omega, t) = \Re[I_{0y}(\omega, t)], \tag{10}$$

A whitening weight has been proposed to address both the problem of room reverberation and the problem of additive noise

$$W(\omega, t) = \sqrt{|P_0(\omega, t)|^2 + \beta(|V_x(\omega, t)|^2 + |V_y(\omega, t)|^2)}. \tag{11}$$

It has also been proven that the weight achieves best localization performance with $\beta = 10^6$. Consequently, we can rewrite the resulting features for the x and y direction as

$$I_x^{(W)}(\omega, t) = \frac{I_x(\omega, t)}{W(\omega, t)}, \tag{12}$$

$$I_y^{(W)}(\omega, t) = \frac{I_y(\omega, t)}{W(\omega, t)}. \tag{13}$$

Enriching the training dataset further improves the robustness, the authors introduced SI features extracted by the four sub-arrays that can be derived from the original structure. Indeed, three microphones are enough to form two orthogonal DMAs in which one microphone is shared. The sub-arrays are groups of three adjacent microphones, in particular we will focus on $M_4 - M_1 - M_2$ (Figure 1), although the same applies for the others. The particle velocity over their direction is given by

$$V_{-\frac{3}{4}\pi, M_1}(\omega, t) = j\sqrt{2}\frac{[P_1(\omega, t) - P_4(\omega, t)]}{\omega \rho d}, \tag{14}$$

$$V_{\frac{3}{4}\pi, M_1}(\omega, t) = j\sqrt{2}\frac{[P_1(\omega, t) - P_2(\omega, t)]}{\omega \rho d}, \tag{15}$$

where M_1 is the shared microphone. Accordingly, the SI features can be written as

$$I^{(W)}_{-\frac{3}{4}\pi, M_1}(\omega, t) = \frac{1}{W_{M_1}(\omega, t)}\Re[P_{M_1}(\omega, t) V^\star_{-\frac{3}{4}\pi, M_1}(\omega, t)] \tag{16}$$

$$I^{(W)}_{\frac{3}{4}\pi, M_1}(\omega, t) = \frac{1}{W_{M_1}(\omega, t)}\Re[P_{M_1}(\omega, t) V^\star_{\frac{3}{4}\pi, M_1}(\omega, t)] \tag{17}$$

where

$$W_{M_1}(\omega, t) = [|P_{M_1}(\omega, t)|^2 + \beta(|V_{\frac{3}{4}\pi, M_1}(\omega, t)|^2 + |V_{-\frac{3}{4}\pi, M_1}(\omega, t)|^2)]^{\frac{1}{2}}, \tag{18}$$

and

$$P_{M_1}(\omega, t) = \frac{P_4(\omega, t) + P_1(\omega, t) + P_2(\omega, t)}{3}. \tag{19}$$

Finally, we define an $M \times N \times C$ matrix Γ to gather all the features together. In particular, M represents the time frames, N the frequency bins and $C = 10$ since the principal array plus the four sub-arrays generate SI features in two direction:

$$\Gamma(:, :, i) = I^{(W)}_{(k)}, \tag{20}$$

where each i from 1 to 10 is respectively associated with the i-th $k = [(x),\ (y),\ (-\frac{3}{4}\pi, M_1),\ (\frac{3}{4}\pi, M_1),\ (-\frac{\pi}{4}, M_2),\ (-\frac{3}{4}\pi, M_2),\ (\frac{\pi}{4}, M_3),\ (-\frac{\pi}{4}, M_3),\ (\frac{\pi}{4}, M_4),\ (\frac{3}{4}\pi, M_4)]$.

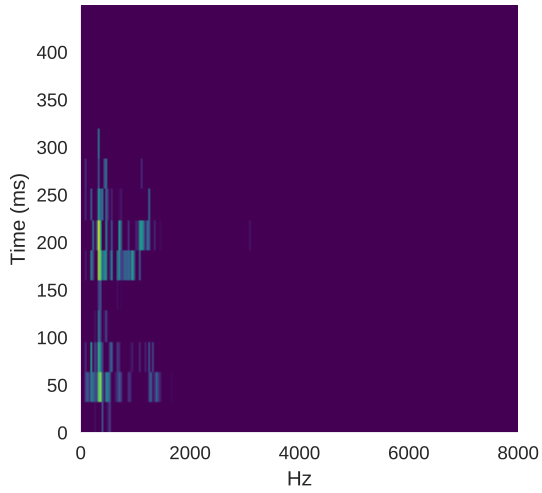


Figure 2: Example of extracted SI features over direction x

3.2. Network Architecture

The CNN is composed of two convolutional layers and three fully-connected layers. It has also the following characteristics:

- The input shape is $14 \times 511 \times 10$
- Each convolutional layer has 64 kernels of size 3×3 with a stride size of 1×1
- A Batch Normalization layer is used after each convolutional layer to improve the stability [16]
- All layers except the output use a ReLU as activation function
- After the last convolutional layer and the first two fully-connected layers, a Dropout procedure with rate 0.5 is applied to avoid overfitting [17]
- The first two fully-connected layers have 512 nodes
- The output layer uses SoftMax as activation function and it has K nodes

We trained two models, one with $K = 12$ that we called `modelRes30` since it classifies with a resolution of 30°, and the other with $K = 36$ called `modelRes10` since it classifies with a resolution of 10°. Hence, the candidate source locations span uniformly the range from -170° to 160° and from -180° to 170° respectively.

4. SIMULATION

In this section, we explain all the details regarding the simulation process. As a starting point, we recreate the training and the validation sets to show that our models achieve the same performance of [12]. From there, we design three other test datasets to test different parameters of the simulation. In particular, we want to study how the two models behave by changing the room volume, the position of the microphone array and the distance of the sources.

4.1. Evaluation Metrics

We utilized the same evaluation metric defined in [12] which is expressed in terms of localization accuracy

$$AR = \frac{N_c}{N_s} \times 100\%, \tag{21}$$

where N_c is the number of locations predicted correctly and N_s is the total number of locations evaluated. A source location is considered correctly evaluated if the candidate source is in the range of $\pm$ the resolution of the model. In other words, adjacent classes to the actual class are considered correct predictions.

4.2. Simulation Setup

In order to generate the datasets, we used `RIR-GENERATOR` [18], an acoustic simulation software based on the image-source method [19], and the speech signals from TIMIT database as sound sources.

For the training dataset, we placed our array (Figure 1) in the center of a rectangular room with dimensions $9.64 \times 7.04 \times 2.95\, m^3$ and at a height of $1.5\ m$. The source is then placed at a distance of $1.5\ m$ from the center of the microphones at the same height. The RIRs are computed with default reflection order and $T60 = 300\ ms$ for 6000 random DOAs uniformly spanning all the directions. Then we selected randomly 300 sentences from the TIMIT database to be convolved with the generated RIRs (sentences are repeated for different DOAs). The same procedure has been followed to generate the validation set with size 1000. While the test datasets for `modelRes30` and `modelRes10` have size 120 and 360 respectively, 10 sentences for each source location.

The STFT is performed with an Hanning window of length 1024 samples and 50% overlap. And since the sentences have a sampling frequency of 16 KHz and have been cut at $500\ ms$, the STFT has shape 14×511[1]. Recalling eqn. (20), we fix the other two dimensions of Γ: $M = 14$ and $N = 511$, which corresponds exactly to the input shape of the CNN.

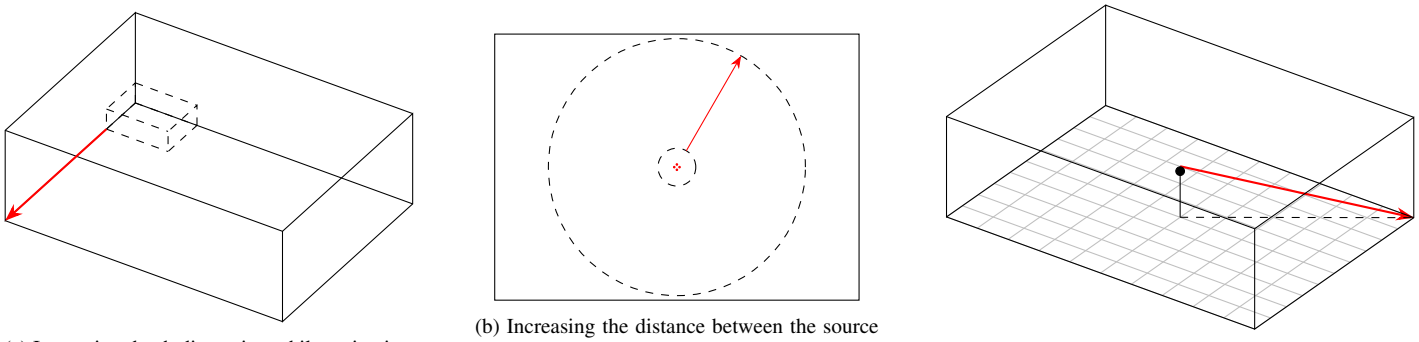
We used Adam [20] as optimizer with an initial learning rate of 10^{-5} and we trained the models for 400 epochs with a patience of 100 epochs on validation accuracy. `modelRes30` reached an accuracy of 100% in testing while `modelRes10` reached 98.9%.

4.3. Custom Test Datasets

Each test dataset is composed of 2000 data points because for each of the 20 steps of the designated parameter we tested 100 different DOAs. An intuitive visualization of the testing ranges is shown in Figure 3, the red arrow corresponds to the x-axis of the result plots in Figure 4. We also mention that the following are a variation of the datasets defined in Section 4.2, the same parameters apply if not otherwise stated.

The first test consists in changing the volume of the room while maintaining the same shape. Hence, we just vary the y dimension of the previously defined room while maintaining constant ratios with respect to the other two dimensions. The y dimension spans the range $[0.16, 15]\ [m]$. And to provide to the simulation some sort of plausibility, we vary also the T60 (the time it takes for the sound pressure level to reduce by 60dB in a reverberant environment) such

[1] Actually, the time and frequency dimensions are transposed with respect to a normal STFT. In that way, the SI features present all the significant values at the beginning, see Figure 2. Furthermore, the first frequency bin is discarded because always equal to zero.

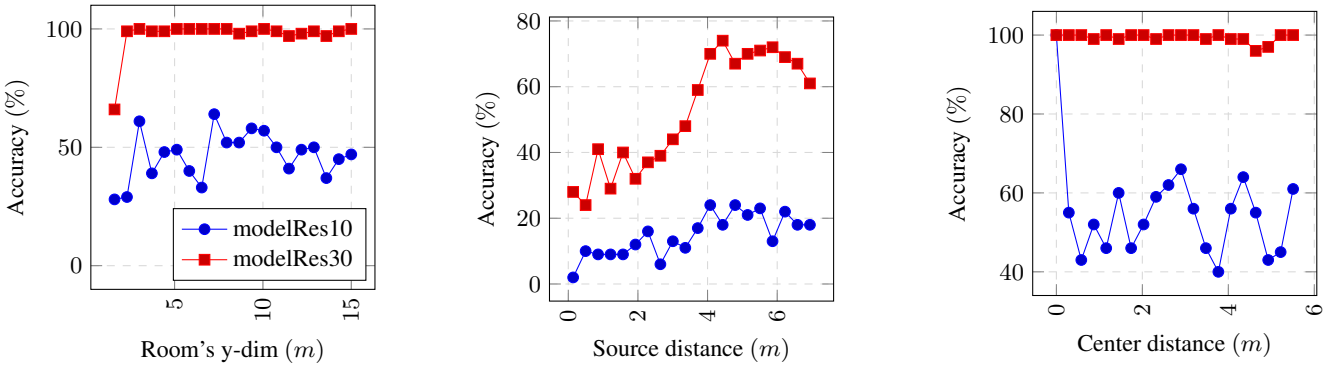


(a) Increasing depth dimension while maintaining the proportions of the room

(b) Increasing the distance between the source and the microphones

(c) Moving the microphones toward the top-right corner at floor height

Figure 3: Experiments visualization. The red arrow represents the testing ranges.



(a) Test 1: room volume

(b) Test 2: source distance

(c) Test 3: microphones position

Figure 4: Performance evaluation for the three experiments

that small rooms are less reverberant while bigger rooms are more reverberant. To achieve this, the T60 follows linearly the volume as stated by Sabine's Formula[2].

The second test consists in increasing the radial distance of the sources from the center of the microphones. The distance spans the range $[0.14, 6.94]$ $[m]$ and consequentially the SI features will vary a lot in magnitude.

The third and last test consists in changing the position of the microphone array by moving it from the center of the room to the top-right corner at floor height. The testing range $[0, 5.5]$ $[m]$ is expressed in terms of distance from the center of the room, where for zero distance we mean the center itself. This test is useful because it breaks the symmetry of the training simulation. Moving the microphones and the sources toward the wall will generate different and new reflection patterns.

5. RESULTS

Observing Figures 4a and 4c we noticed that `modelRes30` presents the best possible performance, while `modelRes10` suffers an overall loss of accuracy that oscillates more in test 3 and less in test 1 around 50%. We think that the main reason lies in the fact that `modelRes30` presents higher intra-class variability in the feature space. Indeed, we used the same dataset to train both models, and since `modelRes30` classifies fewer classes, it has been given 3 times more examples for each class. Another interesting fact is that the accuracy decreases as soon as we are out of the training conditions. This can be clearly seen in the first two steps of `modelRes10` in Figure 4c. A separate question is test 2 (Figure 4b), where the models follow the same trend with `modelRes30` in a more pronounced manner: the more the sources move away from the microphones, the higher the accuracy. This has to be related to the near-field effects of the sources, suggesting that this kind of method is successful only under far-field assumptions.

6. CONCLUSION AND FUTURE WORK

In our work, we tested the validity of the method proposed in [12] and further explored its generalization capabilities over three different simulation parameters. In conclusion, we found that the model generalizes well over the volume of the room and the position of the microphone array as long as we augment the size of the training dataset. Another approach could also be to directly incorporate in the training phase the test dataset we have created, paying attention to the balance of the examples. We also found that the method does not work well in near-field conditions, consequently, a different classification methods should be proposed.

[2] $T60 = 0.161 \frac{V}{A}$, being V the room volume and A the effective absorption area.

7. REFERENCES

[1] A. W. Archer-Boyd, W. M. Whitmer, W. O. Brimijoin, and J. J. Soraghan, "Biomimetic direction of arrival estimation for resolving front-back confusions in hearing aids," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. EL360–EL366, 2015.

[2] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1997, pp. 187–190.

[3] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, "Multi-microphone speaker separation based on deep doa estimation," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[4] H.-Y. Lee, J.-W. Cho, M. Kim, and H.-M. Park, "Dnn-based feature enhancement using doa-constrained ica for robust speech recognition," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1091–1095, 2016.

[5] D. Kurc, V. Mach, K. Orlovsky, and H. Khaddour, "Sound source localization with das beamforming method using small number of microphones," in *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*, 2013, pp. 526–532.

[6] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[7] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[9] J. Dmochowski, J. Benesty, and S. Affes, "Fast steered response power source localization using inverse mapping of relative delays," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 289–292.

[10] A. Marti, M. Cobos, J. J. Lopez, and J. Escolano, "A steered response power iterative method for high-accuracy acoustic source localization," *The Journal of the Acoustical society of America*, vol. 134, no. 4, pp. 2627–2630, 2013.

[11] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[12] N. Liu, H. Chen, K. Songgong, and Y. Li, "Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 1069–1084, 2021.

[13] F. Fahy, *Sound intensity*. CRC Press, 2017.

[14] F. Jacobsen, "Sound intensity," in *Springer Handbook of Acoustics*. Springer, 2014, pp. 1093–1114.

[15] H. Hacıhabiboğlu, "Theoretical analysis of open spherical microphone arrays for acoustic intensity measurements," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 465–476, 2013.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] E. A. Habets, "Room impulse response (rir) generator," 2008. [Online]. Available: https://github.com/ehabets/RIR-Generator

[19] 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.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.