TIME-FREQUENCY MASKING BASED ONLINE SPEECH ENHANCEMENT WITH MULTI-CHANNEL DATA USING CONVOLUTIONAL NEURAL NETWORKS

Soumitro Chakrabarty[1]*, DeLiang Wang*[2] *and Emanuël A. P. Habets*[1]

[1] International Audio Laboratories Erlangen, Germany*
[2] Department of Computer Science and Engineering, and Center for Cognitive & Brain Sciences, The Ohio State University, USA
{soumitro.chakrabarty}@audiolabs-erlangen.de

ABSTRACT

Speech enhancement in noisy and reverberant conditions remains a challenging task. In this work, a time-frequency masking based method for speech enhancement with multi-channel data using convolutional neural networks (CNN) is proposed, where the CNN is trained to estimate the ideal ratio mask by discriminating directional speech source from diffuse or spatially uncorrelated noise. The proposed method operates on, frame-by-frame, the magnitude and phase components of the short-time Fourier transform coefficients of all frequency sub-bands and microphones. The avoidance of temporal context and explicit feature extraction makes the proposed method suitable for online implementation. In contrast to most speech enhancement methods that utilize multi-channel data, the proposed method does not require information about the spatial position of the desired speech source. Through experimental evaluation with both simulated and real data, we show the robustness of the proposed method to unseen acoustic conditions as well as varying noise levels.

Index Terms— convolutional neural networks, speech enhancement, microphone array, masking

1. INTRODUCTION

In modern hands-free communication systems, extraction of a desired speech signal in noisy and reverberant environments is an important task. With the advent of multiple microphones on modern devices, microphone array processing techniques for spatial filtering methods have become an attractive solution to the task. Spatial filtering techniques [1] rely on utilizing spatial information regarding the sound scene within an estimation framework, to extract the desired speech signal and suppress the undesired signal components. Though with accurate information regarding the sound scene these methods can perform well, in practice, information such as spatial position of the speech source, second order statistics (SOS) of the desired and the undesired signal components etc. is required. Such spatial information generally needs to be estimated, which itself is a challenging task in adverse acoustic conditions.

Following the recent success of deep learning based approaches for different signal processing tasks, methods have been proposed for performing speech enhancement with multi-channel data. Based on the estimation target, two main types of approaches have been proposed: i) using deep neural networks (DNNs) to estimate beamformer weights [2–5]; ii) using DNNs to estimate time-frequency (T-F) masks [6–9].

Methods of the first type train a DNN to estimate the real and imaginary parts of the weights of a beamformer. In [2], a DNN is used to estimate the weights of a delay-and-sum beamformer, whereas in [5], a recurrent neural network (RNN) estimates the weights of a filter-and-sum beamformer. Though both methods showed improvement in terms of speech recognition performance, both the target beamformers considered in these works are data-independent beamformers which are known to be limited in terms of their overall speech enhancement ability, especially in noisy and reverberant conditions [1].

The work presented in this paper is related to the second type of methods that relies on multi-channel data to estimate T-F masks, which can either be used to estimate the SOS of the signal components for spatial filtering or directly utilized as a real-valued gain to obtain the desired signal. In [7], only spatial features from the microphone signals are used for training a DNN to estimate a T-F mask. In [6, 9], a binaural setup is considered where both spatial and spectral features are computed to train a network to obtain an estimate of the ideal binary mask (IBM) or the ideal ratio mask (IRM) [10], which is then applied as a real-valued gain to obtain the desired signal. All the above mentioned methods involve explicit feature extraction steps, which can lead to considerable computational cost, especially when the number of microphones is high. Also the spatial position of the desired speech source was considered fixed [6] or known [7, 9]. In [8], neural networks with shared weights for each channel were used to estimate an IBM, which was then utilized to compute the power spectral density (PSD) of the desired and noise components. As each channel was treated separately, no spatial information was exploited for mask estimation.

In this work, we propose a T-F masking based approach for speech enhancement where a convolutional neural network (CNN) is trained to discriminate between the spatial characteristics of the desired directional speech source and diffuse or uncorrelated noise components in order to estimate the IRM for extracting the desired speech source at the output. Instead of an explicit feature extraction step, the magnitude and phase of the short-time Fourier transform (STFT) coefficients of all the frequency sub-bands and microphones, for a single STFT time frame, are directly provided as input to the system. Without an explicit spatial feature extraction step or any temporal context in the input representation, makes the proposed method suitable for online speech enhancement. Additionally, the proposed method is designed to be independent of the spatial position of the desired speech source. In this paper, we consider an acoustic scenario with a single desired speech source in the pres-

* A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

978-1-5386-8151-0/18/$31.00

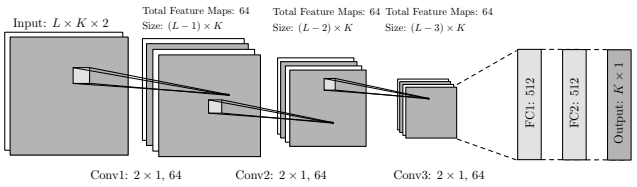

Fig. 1. Proposed CNN architecture.

ence of diffuse babble noise as well as microphone self-noise, and evaluate the performance of the proposed method in unseen acoustic conditions with both simulated and real data.

2. PROBLEM FORMULATION

Let us consider an array of L microphones that captures the reverberant speech signal along with noise. In the STFT domain, the vector of received signals, $\mathbf{y}(n,k) = [Y(n,k,1), \ldots, Y(n,k,L)]^{\mathrm{T}}$ at time frame n and frequency bin k is given by

$$\mathbf{y}(n,k) = \mathbf{x}_d(n,k) + \mathbf{x}_r(n,k) + \mathbf{v}_d(n,k) + \mathbf{v}(n,k), \quad (1)$$

where $\mathbf{x}_d(n,k)$ and $\mathbf{x}_r(n,k)$ denote the direct and the reverberant part of the speech signal, respectively. We consider two different kinds of noise components present in the sound scene; diffuse babble noise, denoted by $\mathbf{v}_d(n,k)$, and spatially uncorrelated microphone self-noise, denoted by $\mathbf{v}(n,k)$. Given the recorded microphone signals, the aim in this work is to extract the direct component of the speech signal at a reference microphone $X_d(n,k,1)$.

To obtain an estimate of the desired signal, we propose a time-frequency (T-F) masking based approach, where we want to estimate the positive and real-valued ideal ratio mask (IRM) [10], given by

$$M_{\mathrm{IRM}}(n,k) = \frac{|X_d(n,k,1)|}{|Y(n,k,1)|}, \quad (2)$$

where $Y(n,k,1)$ denotes the signal recorded at the first microphone. With the mask, an estimate of the desired signal is given by

$$\widehat{X}_d(n,k,1) = M_{\mathrm{IRM}}(n,k) \ \cdot \ Y(n,k,1). \quad (3)$$

Inverse STFT using the estimated magnitude and the noisy and reverberant phase is performed to obtain the desired signal waveform. In the following sections, we describe the proposed method to obtain an estimate of the IRM using the multi-channel data and a CNN.

3. PROPOSED SYSTEM

The aim of the proposed method is to obtain an accurate estimate of the IRM for all the frequency bins of a single time frame, given the input feature representation of the corresponding time frame. Please note that no temporal context is utilized in this work.

3.1. Input feature representation

In the proposed system, the input corresponds to a single time frame of the STFT representation of the microphone signals. The observed signal can be expressed as

$$Y_l(n,k) = A_l(n,k)e^{j\phi_l(n,k)}, \quad (4)$$

where $A_l(n,k)$ represents the magnitude component and $\phi_l(n,k)$ denotes the phase component of the STFT coefficient of the received signal at the l-th microphone. Rather than performing explicit feature extraction, we directly provide the magnitude and the phase of the STFT coefficients of the received signals as input to the system. The idea is to have the system learn to differentiate between the directional speech source and the diffuse/uncorrelated noise components during training in order to estimate the IRM.

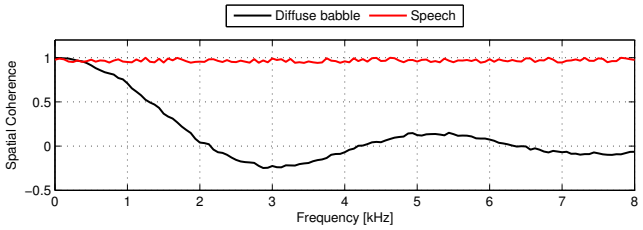

Fig. 2. Spatial coherence of speech vs babble noise.

The input feature for the n-th time frame is formed by arranging $A_l(n,k)$ and $\phi_m(n,k)$ for each time-frequency bin (n,k) and each microphone l into a three dimensional tensor of size $L \times K \times 2$, which we call the *STFT map*, where $K = N_f/2 + 1$ is the total number of frequency bins, upto the Nyquist frequency, at each time frame and L is the total number of microphones in the array.

3.2. CNN for mask estimation

Given the input STFT map for each time frame, $\mathbf{\Phi}(n)$, we employ a CNN to learn to differentiate between the spatial characteristics of the speech and noise sources, and obtain an accurate estimate of the IRM for all the frequency bins corresponding to that time frame, $\widehat{\mathbf{M}}(n) = \left[\widehat{M}(n,1), \ldots, \widehat{M}(n,K)\right]$.

In Fig. 1, we show the CNN architecture employed in this work, which is similar to the architecture proposed in our previous work for the task of DOA estimation [11]. In the convolution layers (Conv layers in Fig. 1), small filters of size 2×1 are applied to the input. The choice of such small filters is motivated by the discriminative feature that the CNN can learn to differentiate between sound sources with different spatial characteristics.

To illustrate this point, in Fig. 2, for a microphone pair separated by 8 cm, the difference in the spatial coherence [12] between the speech source and the diffuse babble noise, assuming an ideal isotropic spherically diffuse noise field is shown. It can be seen that while the speech signal is highly coherent across the complete spectrum, the diffuse noise is coherent only in the very low frequency region. For the CNN to effectively learn this frequency-dependent discriminative feature, small filters are applied that can potentially learn from the correlation between the STFT coefficients of neighboring microphones for each of the frequency sub-bands separately. These learned local structures are then eventually combined by the fully connected layers (FC layers in Fig. 1) for the final mask estimation.

For both the convolution as well as the fully connected layers, in this work, we use the rectified linear units (ReLU) activation function [13]. In the final layer of the network, as the IRM for each T-F bin lies between 0 and 1, we use K sigmoid units to estimate the IRM for a given time frame. Though the IRMs for all the frequency bins of a single time frame are estimated simultaneously, the optimization is done separately for each sigmoid output corresponding to each frequency bin. For the optimization of the weights of the CNN, we use the L_2 or mean-square error (MSE) loss function,

Simulated training data	
Signal	Speech signals from TIMIT
Room size	R1: (6 × 6), R2: (5 × 4), R3: (10 × 6), R4: (8 × 3), R5: (8 × 5)
Array positions in room	7 different positions in each room
Source-array distance	1 m and 2 m for each position
RT_{60}	R1: 0.3 s, R2: 0.2 s, R3: 0.8 s, R4: 0.4 s, R5: 0.6 s
SNR	Diffuse babble: -6 to 6 dB, Spatially white: 5 to 20 dB

(a)

Simulated test data	
Signal	Speech signals from LIBRI
Room size	Room 1: (4 × 7) , Room 2: (9 × 4)
Array positions	3 arbitrary positions in each room
Source-array dist.	1.7 m for both rooms
$RT_{60}(s)$	Room 1: 0.38 , Room 2: 0.70

(b)

Table 1. (a) Configuration for training data generation. All rooms are 2.7 m high. (b) Configuration for simulated test data generation. All rooms are 3 m high.

Rooms	Room 1									Room 2								
SNR	−6 dB			0 dB			6 dB			−6 dB			0 dB			6 dB		
Measure	Δ SNR	Δ P	Δ ST	Δ SNR	Δ P	Δ ST	Δ SNR	Δ P	Δ ST	Δ SNR	Δ P	Δ ST	Δ SNR	Δ P	Δ ST	Δ SNR	Δ P	Δ ST
DSB	8.6	0.21	0.08	9.2	0.23	0.09	9.7	0.30	0.08	8.5	0.10	0.08	9.2	0.22	0.09	9.1	0.25	0.08
RSD	**9.1**	0.25	0.14	**9.8**	0.30	0.12	**9.9**	0.34	0.10	**8.8**	0.13	0.12	**9.5**	0.26	0.11	**9.4**	0.30	0.09
Proposed	3.8	**0.30**	**0.16**	4.3	**0.43**	**0.14**	4.8	**0.45**	**0.11**	3.8	**0.21**	**0.15**	4.2	**0.36**	**0.13**	4.8	**0.39**	**0.10**
Oracle	19.4	1.48	0.40	19.5	1.56	0.31	19.8	1.53	0.25	19.3	1.38	0.42	19.2	1.54	0.31	19.6	1.53	0.24

Table 2. Results for two different simulated acoustic conditions with varying levels of diffuse babble noise. The best performing methods are shown in bold. Δ P and Δ ST denote PESQ and STOI score improvement, respectively.

given by

$$\mathcal{L} = \left(M_{\text{IRM}}(n,k) - \widehat{M}(n,k) \right)^2 . \tag{5}$$

The different hyper-parameters of the proposed architecture in Fig. 1 were chosen by using a validation data set.

The CNN is trained using a training data set $\{\{\mathbf{\Phi}(n), \mathbf{M}_{\text{IRM}}(n)\} \mid n = 1, \ldots, N\}$, where N denotes the total number of STFT time frames in the training set, and $\mathbf{M}_{\text{IRM}}(n)$ denotes the IRM vector for the n-th time frame. Details regarding the preparation of the training data set are given in Section 4.1.

In the test phase, the test signals are first transformed into the STFT domain using the same parameters used during training. Following this, the STFT map for each time frame of the test signals is given as input to the CNN, and the CNN provides an estimate of the IRM for each frequency bin in each time frame. Finally, an estimate of the desired signal is obtained using (3).

4. PERFORMANCE EVALUATION

4.1. Training details

We consider a uniform linear array (ULA) with $L = 4$ microphones with inter-microphone distance of 8 cm, and the input signals, with sampling frequency of $\text{F}_s = 16$ kHz, are transformed to the STFT domain using a DFT length of 256, with 50% overlap, resulting in $K = 129$. The room impulse responses (RIRs) required to simulate different acoustic conditions are generated using the RIR generator [14].

The configurations for generating the training data are given in Table 1a. For training, we used 370 randomly chosen speech utterances from the TIMIT dataset, each 2 s long. For the proposed method to be independent of the spatial position of the desired speech source, for each array position and source-array distance considered in the training conditions, the whole angular range of the ULA was discretized with a 5° resolution, to get 37 different angular positions of the speech source. For each of these positions, 10 speech signals were convolved with the simulated RIRs corresponding to that specific setup. Then, spatially uncorrelated Gaussian noise was added to the training data with randomly chosen SNRs between 5 and 20 dB. Additionally, diffuse babble noise, with randomly chosen SNRs between -6 and 6 dB was also added. A 40 s long sample of multi-channel diffuse babble noise was generated using the acoustic noise field generator [12], assuming an isotropic spherically diffuse noise field. The generated babble noise was divided into 20 segments, each of length 2 s. The first 10 segments were used for training, and the rest were used for test. In total, the training data consisted of around 8.3 million time frames, which was approximately 18 hours worth data.

The CNN was trained using the Adam gradient-based optimizer [15], with mini-batches of 512 time frames and a learning rate of 0.001. During training, at the end of the three convolution layers and after each fully connected layer, a dropout procedure [16] with a rate of 0.5 was used to avoid overfitting. All the implementations were done in Keras [17].

4.2. Baselines and performance measures

The performance of the proposed method is compared to two traditional beamformers, delay-and-sum beamformer (DSB) and the robust super-directive beamformer (RSD) [1]. To compute the relative transfer function (RTF) for both beamformers, first a white noise signal is convolved with the direct part of the RIR. Then with the STFT representation of this signal, for each frequency sub-band, the PSD matrix is computed with a long-term average. From this PSD matrix, depending on the reference microphone, the corresponding column, normalized with respect to the signal power at the reference microphone, is used as the propagation/steering vector. Additionally, for the RSD beamformer, the theoretical coherence matrix for a spherically isotropic diffuse noise field is used as the undesired signal PSD, with diagonal loading applied to avoid singularity at low frequencies. Please note that the beamforming methods used for comparison are their corresponding ideal versions with perfect in-

RT_{60}	0.160 s						0.360 s						0.610 s					
Distances	1 m			2 m			1 m			2 m			1 m			2 m		
Measure	Δ SNR	Δ P	Δ ST	Δ SNR	Δ P	Δ ST	Δ SNR	Δ P	Δ ST	Δ SNR	Δ P	Δ ST	Δ SNR	Δ P	Δ ST	Δ SNR	Δ P	Δ ST
DSB	8.5	0.22	0.11	8.2	0.20	0.11	8.5	0.21	0.10	8.2	0.14	0.09	8.5	0.20	0.09	8.3	0.17	0.07
RSD	**8.7**	0.25	0.13	**8.3**	0.27	0.12	**8.6**	0.29	0.11	**8.5**	0.19	0.10	**8.6**	0.23	**0.12**	**8.4**	0.18	0.09
Proposed	4.3	**0.42**	**0.14**	3.6	**0.32**	**0.13**	3.8	**0.39**	**0.14**	3.8	**0.21**	**0.13**	3.3	**0.27**	**0.12**	3.2	**0.21**	**0.11**
Oracle	20.1	1.41	0.32	19.7	1.40	0.33	20.0	1.42	0.33	19.6	1.34	0.37	20.2	1.46	0.36	19.7	1.37	0.41

Table 3. Results for different distances and reverberation times in real acoustic conditions. The best performing methods are shown in bold.

formation. In practice, these methods would require an estimate of the direction-of-arrival (DOA) of the source or the RTF.

The compared methods are evaluated in terms of frequency weighted segmental signal-to-noise ratio improvement (ΔSNR) [18], PESQ improvement (ΔP) [19] and improvement in terms of short-term objective intelligibility (ΔST) score [20].

4.3. Test results

4.3.1. Simulated RIRs

For test in simulated conditions, we used a total of 185 different speech utterances from the LibriSpeech ASR corpus [21], 5 utterances per angular position for each array setup. Each utterance was 2 s long. The acoustic conditions for test are shown in Table 1b.

The results in simulated acoustic conditions are shown in Table 2. The performance of the methods was evaluated for three different input SNRs for the babble noise, while the self-noise input SNR was 10 dB for all the cases. In addition to the three competing methods, objective results for the oracle IRM applied to extract the desired speech source is also shown (Oracle). The shown results for each input SNR of babble noise was averaged over all the different angular positions of the desired source for a given array position, as well as the different array positions considered for each room.

From the results, it can be seen that the proposed method achieves significant enhancement, in terms of STOI score and PESQ improvement, even in highly reverberant conditions (Room 2). Also, both the proposed method and the RSD beamformer achieve the maximum improvement in objective intelligibility (ΔST) when the level of diffuse babble noise is high, whereas for the DSB, the performance is constant for the different SNRs of babble, since it is mainly designed to reduce uncorrelated noise which is constant for all conditions. In terms of SNR improvement, the best performance is always achieved by the RSD beamformer, with the performance of the DSB being slightly inferior. The proposed method demonstrates significantly less SNR improvement.

Overall, in terms of STOI and PESQ improvement, the proposed method achieves the best performance, however from the results provided for the oracle IRM (Oracle), it can be seen that there remains a significant scope for improvement in results of the proposed method.

4.3.2. Measured RIRs

To further investigate the robustness of the proposed method to different acoustic conditions, we evaluated the performance of the methods with measured RIRs. It should be noted that the proposed method was only trained with simulated data. For this evaluation, we used the Multichannel Impulse Response Database from Bar-Ilan university [22]. The database consists of measured RIRs with sources placed on a grid of $[0^\circ, 180^\circ]$, in steps of 15°, at distances of 1 m and 2 m from the array. For our experiment, we chose the four middle microphones from the array setup with 8 cm inter-microphone distance [22] to have a similar geometric setup in both simulated and real conditions. The test data was generated by convolving 5 different speech utterances, randomly chosen from the LibriSpeech corpus, with the measured RIRs for each of the thirteen discrete angles in the database.

The results for different reverberation times and distances are shown in Table 3. For all the different acoustic conditions, spatially white noise and diffuse babble noise were added to the test signal to obtain an average segmental SNR of 10 dB and 0 dB, respectively. The results shown here are averaged over all the thirteen angles for each acoustic setup.

From the results, we see that that the proposed method, though trained with only simulated data, manages to achieve considerable enhancement across all different real acoustic conditions. Similar to the results in simulated acoustic conditions, the proposed method always achieves the best performance in terms of STOI score and PESQ improvement, but has a significantly lower SNR improvement.

It should be noted that the proposed approach is essentially a single channel enhancement method that utilizes the multi-channel data to estimate the mask. Due to this, as well as the reconstruction of the signal with noisy and reverberant phase, we notice some amount of distortion introduced in the enhanced signal for all the investigated scenarios, which possibly leads to the lower performance in terms of Δ SNR. To further illustrate this, audio examples are provided on our website[1] for both simulated as well as real acoustic conditions.

5. CONCLUSION

A T-F masking based speech enhancement method using a CNN was proposed, and without explicit feature extraction and temporal context it is suitable for online implementations. Based only on spatial features, the proposed method showed significant enhancement performance in both simulated and real acoustic conditions. However, for both cases, we observed a significant gap in performance compared to what could be achieved with the oracle IRM.

In future work, we would like to investigate the improvement in performance that can be achieved by including temporal context as well as spectral features. Additionally, we would also like to extend this work to deal with directional noise sources. Since the input feature representation and the CNN architecture used here are similar to what was used in our previous work on DOA estimation [11], it would also be interesting to investigate whether transfer learning leads to improvement in performance for either task.

[1] www.audiolabs-erlangen.de/resources/2018-IWAENC-CNNSpEnh

6. REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[2] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5745–5749.

[3] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Interspeech 2016*, 2016, pp. 1976–1980.

[4] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 965–979, May 2017.

[5] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 271–275.

[6] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2112–2121, Dec 2014.

[7] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time-frequency masks from spatial features," *Speech Communication*, vol. 68, no. C, pp. 97–106, Apr. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2015.01.006

[8] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 196–200.

[9] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1075–1084, May 2017.

[10] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec 2014.

[11] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2017.

[12] E. A. P. Habets and S. Gannot, "Generating Sensor Signals in Isotropic Noise Fields," *Journal Acoust. Soc. of America*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

[13] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Intl. Conf. on Machine Learning (ICML)*, J. Frnkranz and T. Joachims, Eds. Omnipress, 2010, pp. 807–814. [Online]. Available: http://www.icml2010.org/papers/432.pdf

[14] E. A. P. Habets. (2016) Room Impulse Response (RIR) generator. [Online]. Available: https://github.com/ehabets/RIR-Generator

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, 2014.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, Jan. 2014.

[17] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[18] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Interspeech Conf.*, 2006, pp. 1447–1450.

[19] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[22] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Sept 2014, pp. 313–317.