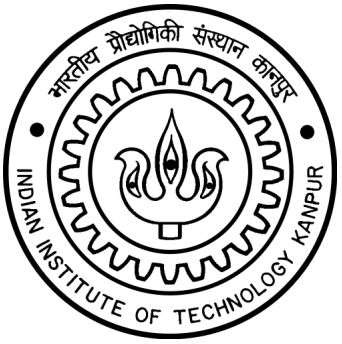

Indian Institute of Technology, Kanpur

EE392: Undergraduate Project

Learning based approach for Multi-channel Speech Enhancement

Paridhi Maheshwari (14448)
Department of Electrical Engineering
Indian Institute of Technology, Kanpur

Supervisor:
Rajesh M. Hegde
Department of Electrical Engineering
Indian Institute of Technology, Kanpur

Abstract

We propose a learning based approach to steering vector estimation for better acoustic beamforming in presence of noisy and reverberant environment. A recurrent convolutional neural network (RCNN) is used to learn the mapping from normalized observed acoustic signals to corresponding steering vector. Earlier works either used parametric mask estimation approaches to steering vector estimation or a learning approach that did not exploit the multi-channel information simultaneously. With RCNN, we aim to avoid processing of individual frequencies like in a parametric mask estimation approach and at the same time exploit the multi-channel information to obtain an accurate steering vector that results in better speech enhancement as demonstrated by R2 score and mean square error. We also demonstrate the results of acoustic beamforming with the estimated steering vector.

1 Introduction

Speech enhancement in noisy environments is of paramount importance in various speech processing tasks including Automatic Speech Recognition (ASR) [12], hearing aids design [22] and Speech coding [13]. Various methodologies have been developed for this purpose. Spectral subtraction [1] generates an estimate of the noise spectrum and subtracts it from the noisy speech spectrum to obtain an estimate of clean speech spectrum. Other conventional methodologies include MMSE based spectral amplitude estimator [4], Wiener filtering [14] and non-negative matrix factorization [23]. However, most of these unsupervised techniques rely on the statistical properties of speech and noise signals, or on the additive nature of noise, and in realistic scenarios, the enhanced speech signals are often characterized by an unwanted component known as "musical noise" [15]. The advent of deep learning has indicated improved performances in both speech enhancement as well as speech separation. This is due to the fact that the mapping between speech and noise can be better modelled using non-linear transformations. One of the methods involve deep recurrent neural networks such as long short-term memory (LSTM) [5]. Modelling the temporal-spectral structures of a spectrogram using convolutional neural networks (CNN) have also been investigated [7].

In case of far-field speech enhancement tasks, it has been observed that using multiple microphones for speech acquisition and leveraging the properties of multi-channel data is also beneficial. Acoustic beamforming for multi-channel input in one such technique which has shown substantial gains in performance. Given a steering vector, a beamformer attenuates signals from every direction other than the target direction specified by the steering vector. It is evident that accurate estimation of steering vector is important. Previous techniques were based on finding the Direction Of Arrival (DOA) of the speaker and obtaining the steering vector from the DOA estimate, when the array geometry is known. Some of the popular techniques for DOA estimation include the Multiple signal classification (MUSIC) [16], Maximum likelihood based modelling [18] and Generalized cross correlation based approaches such

as steered response power with phase transform (SRP-PHAT) [2]. But traditional DOA estimates rely on the inaccurate plane wave assumption and although they give reasonable results for anechoic space, they fail in case of reverberant environments.

This caused a shift from the conventional beamforming approaches to data-driven beamforming in several recent studies [9]. Predictions of spectral masks for speech and noise are utilized to estimate the speech covariance matrix, which further produce the beamforming coefficients.

In this work, we propose a learning based approach to RIR estimation followed by data-driven Minimum Variance Distortionless Response (MVDR) beamforming to obtain a cleaner version of the noisy signal. The rest of the document is organized as follows: Section 2 includes the prior work in this field, Section 3 and 4 describe the system architecture in full detail, Section 5 talks about the different types of experiments that were run along with their results.

2 Related Work

There exist two classes of studies for speech enhancement using spectral mask estimation. The first class includes model based techniques where signals are parametrized by various generative models such as Watson mixture model [21, 17] or Complex Gaussian Mixture Model (CGMM) [9]. Our approach offers two advantages over these techniques: (i) We make no assumptions about the signal structure. The network is trained to capture speech characteristics without any additional information. (ii) Unlike parametric mask estimation tasks, different frequencies are processed simultaneously using the same architecture. This reduces the computational complexity.

The other techniques involve learning based approaches to steering vector estimation. Heymann et al. [8] experimented with feed-forward and a bi-directional Long Short-Term Memory (BLSTM) networks to obtain binary masks for noise and target signal. They work on a per channel basis and are therefore, indifferent to the array configuration. Another recent approach by Erdogan et al. [6] proposed a LSTM network for predicting single channel enhancement masks and combining them for multi-channel beamforming. A stark difference between the above approaches and our approach is that we deal with multi-channel data simultaneously and also, try to exploit the inter-channel information (for a given microphone array).

3 Signal Model

In the time domain, clean speech signal is signal is convolved with the room impulse response (RIR) and then, noise is added. In the Short-time Fourier transform (STFT) domain, where convolution reduces to multiplication, this can be modelled as

$$\mathbf{y}_{f,t,m} = \mathbf{h}_{f,m} \cdot \mathbf{x}_{f,t} + \mathbf{n}_{f,t,m}$$

– $\mathbf{y}_{f,t,m}$ is the STFT of the signal received at the m^{th} microphone.

– $\mathbf{x}_{f,t}$ is the STFT of the clean speech signal.

– $\mathbf{h}_{f,m}$ is the Fourier transform of the RIR of m^{th} microphone.

– $\mathbf{n}_{f,t,m}$ is the STFT of the noise added to the m^{th} microphone.

Note that $\mathbf{x}_{f,t}$ is not dependent on the microphone index as it is the clean speech signal common for all $m = 1, 2 \cdots M$. Also, $\mathbf{h}_{f,m}$ is not a function of t as the impulse response is static, unlike the dynamic speech signals. The complete signal at the (f, t) bin is given by

$$\mathbf{y}_{f,t} = [\mathbf{y}_{f,t,1} \;\; \mathbf{y}_{f,t,2} \;\cdots\; \mathbf{y}_{f,t,M}]^T$$

.

4 Proposed Algorithm

4.1 Feature Extraction

It was observed, through experimental investigations, that the magnitude variance of $\mathbf{y}_{f,t,m}$ as well as $\mathbf{h}_{f,m}$ is much smaller when compared to the variance of phase across different frequencies and channels. Thus, only the phase component of the STFT signals should suffice to find a mapping between the two sets of data. This hypothesis was also supported by Chakrabarty et al. [3] where only the phase component was exploited to find an accurate DOA estimate.

For a fixed microphone array, the relative phase difference with respect to a reference microphone is of primary importance to capture the phase characteristics of RIR. In order to maintain the cyclic property of phase, sin and cosine of the phase difference is used as input.

Essentially, at a given time-frequency bin (f, t), the input features are of the form

$$\begin{bmatrix} \cos(\angle\mathbf{y}_{f,t,2} - \angle\mathbf{y}_{f,t,1}) \\ \sin(\angle\mathbf{y}_{f,t,2} - \angle\mathbf{y}_{f,t,1}) \\ \vdots \\ \cos(\angle\mathbf{y}_{f,t,M} - \angle\mathbf{y}_{f,t,1}) \\ \sin(\angle\mathbf{y}_{f,t,M} - \angle\mathbf{y}_{f,t,1}) \end{bmatrix}$$

The spectrogram is computed for all the channels using a N-point Discrete Fourier Transform (DFT) and a Hammming window function. Since the DFT of real signals is symmetric, only the first $N/2 + 1$ frequencies are unique. The final dimension of the network input is $L \times (2M - 2) \times (\frac{N}{2} + 1)$ where L is the number of time frames.

The output labels, which are independent of the temporal dimension, have the form

$$\begin{bmatrix} \cos(\angle \mathbf{h}_{f,2} - \angle \mathbf{h}_{f,1}) \\ \sin(\angle \mathbf{h}_{f,2} - \angle \mathbf{h}_{f,1}) \\ \vdots \\ \cos(\angle \mathbf{h}_{f,M} - \angle \mathbf{h}_{f,1}) \\ \sin(\angle \mathbf{h}_{f,M} - \angle \mathbf{h}_{f,1}) \end{bmatrix}$$

Similar to the input features, the dimension of the labels is $(2M-2) \times (\frac{N}{2}+1)$. Essentially, the network is a mapping from the 3D input features to the 2D output labels.

4.2 Learning based RIR Estimation

Convolutional Neural Network (CNN) layers are used to extract local shift-invariant properties of the input 3 dimensional data. The temporal dimensions are unaffected in all the CNN layers. Effectively, the 2-D data corresponding to every frame is processed separately. Every CNN layer is provided a rectified linear unit (ReLU) activation and the output is reshaped such that the depth of every layer is incorporated into the third dimension. This ensures that the input to next layer also has 3 dimensions. Also, the data size is reduced with each layer and the final CNN layer reduces the 2-D feature matrix for each time frame into a single feature vector. Each time frame has a distinct feature vector, and therefore, the 3-D data is finally reduced to 2 dimensions after all the CNN layers.

As the the noise characteristics can only be exploited in temporal space, the data is further fed into Recurrent Neural Network (RNN) layers. This is used to learn information about the noise statistics. In more definite terms, the layers are Gated Recurrent Units (GRU) with hyperbolic tangent (tanh) activation.

Lastly, a dense connected layer translates the data into the final output. Since the required output is of regression, a linear activation is used.

Training Scheme: All the layers were trained from scratch. For back-propagation, the Adam gradient descent algorithm [10] was used with a batch size of 32. In order to avoid over-fitting, the early stopping regularization is incorporated wherein the training is stopped if the validation loss does not decrease over a patience of 20 epochs.

Evaluation: We use the Mean Squared Error (MSE) and R^2 Coefficient of determination to quantify the learning. R^2 is a statistical measure of how well the regression output approximates the real data points.

4.3 Data-driven Beamforming

Minimum variance distortionless response (MVDR) beamformer attempts to null signals coming from any direction other than the the steering vector. If the beamformer is represented by a linear filter $\mathbf{w}_f$ and steering vector by $\mathbf{r}_f$, the optimization problem can be expressed as:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathbf{w}^H \mathbf{S}_{\mathbf{yy}} \mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^H \mathbf{r}_f = 1 \end{aligned}$$

The closed form solution to this problem is given by

$$\mathbf{w}^H = \frac{\mathbf{r}_f^H \mathbf{S}_{\mathbf{yy}}^{-1}}{\mathbf{r}_f^H \mathbf{S}_{\mathbf{yy}}^{-1} \mathbf{r}_f}$$

And, the spectrum of the enhanced signal is given by

$$\hat{\mathbf{s}}_{f,t} = \mathbf{w}_f^H \mathbf{y}_{f,t}$$

5 Experiments and Results

5.1 RIR Estimation for certain frequencies

Layer (type)	Output Shape	# Parameters
Time Distributed	(15000, 30, 13, 1, 15)	315
Reshape	(15000, 30, 13, 15, 1)	0
Time Distributed	(15000, 30, 10, 1, 20)	1220
Reshape	(15000, 30, 10, 20, 1)	0
Time Distributed	(15000, 30, 6, 1, 25)	2525
Reshape	(15000, 30, 6, 25, 1)	0
Time Distributed	(15000, 30, 3, 1, 40)	4,040
Reshape	(15000, 30, 3, 40, 1)	0
Time Distributed	(15000, 30, 1, 1, 140)	16,940
Reshape	(15000, 30, 140)	0
GRU	(15000, 30, 100)	72,300
GRU	(15000, 75)	39,600
Dense	(15000, 140)	10,640
Reshape	(15000, 14, 10)	0
Total		1,47,580

Table 1: Brief summary of the model architecture

A smaller architecture was first applied to a small set of frequencies, to validate the scalability of the approach. The model was trained on $15K$ recording samples, each consisting of 30 time frames and the frequency set $101 - 110$. A 2048 point DFT

was computed in all the experimentation. For validation, a different set of $1K$ recordings were used with similar specifications. The net training input to the network is of size $(15000, 30, 14, 10, 1)$ and validation set is of $(1000, 30, 14, 10, 1)$ dimension.

Each epoch took approximately 90 seconds and convergence was achieved in 11 epochs. The best model yielded a MSE of 0.2840 and R^2 score of 40.21% on the validation data.

The predictions and labels for a single recording have been plotted in Figure 1. Although there are a few distortions, the model roughly captures the entire pattern structure. We believe that increasing the number of frequencies will provide more inter-frequency information and the network will be able to learn more intricate patterns as well.

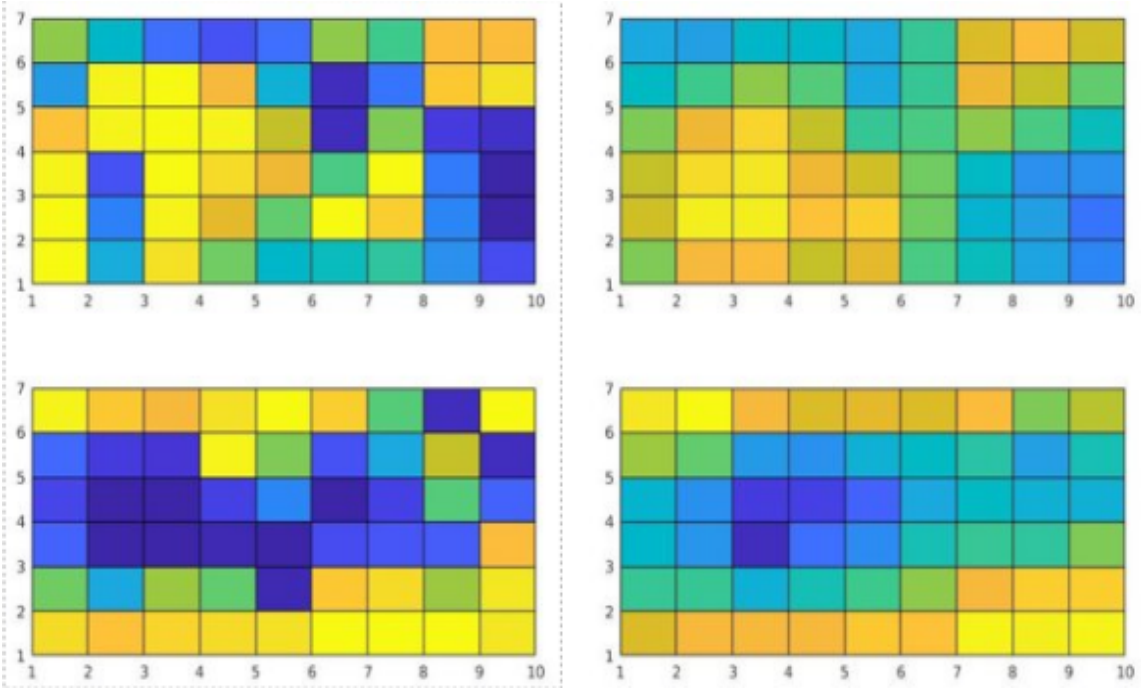

Figure 1: Results of trained model for one testing recording (a) Cosine of ideal labels (b) Cosine of predicted labels (c) Sine of ideal labels (d) Sine of predicted labels

5.2 Comparison with Model based approach

The above experimental setup was compared to a model based approach proposed by Higuchi et al. [9]. They model the STFT signals of speech and noise using a Complex Gaussian Mixture Model (CGMM) and the parameters are estimated by an iterative Expectation-Minimization (EM) algorithm. Since the computation is done on a per frequency basis, it is possible to duplicate the setup of the previous experiment.

Unlike the proposed algorithm, the CGMM method predicts both magnitude and phase. For the same frequency range, it was observed that the variance of magnitude was 10^{-3} times that of phase. This justifies our hypothesis of neglecting the magnitude. For identical recordings, the algorithm suffered a MSE of 1.0120 for the cos and sin of the phase of the predicted steering vector. Our algorithm, on the other hand, incurred a loss of 0.2840. It is evident that the proposed algorithm outperforms the CGMM modelling approach.

5.3 Extension to Speech Enhancement

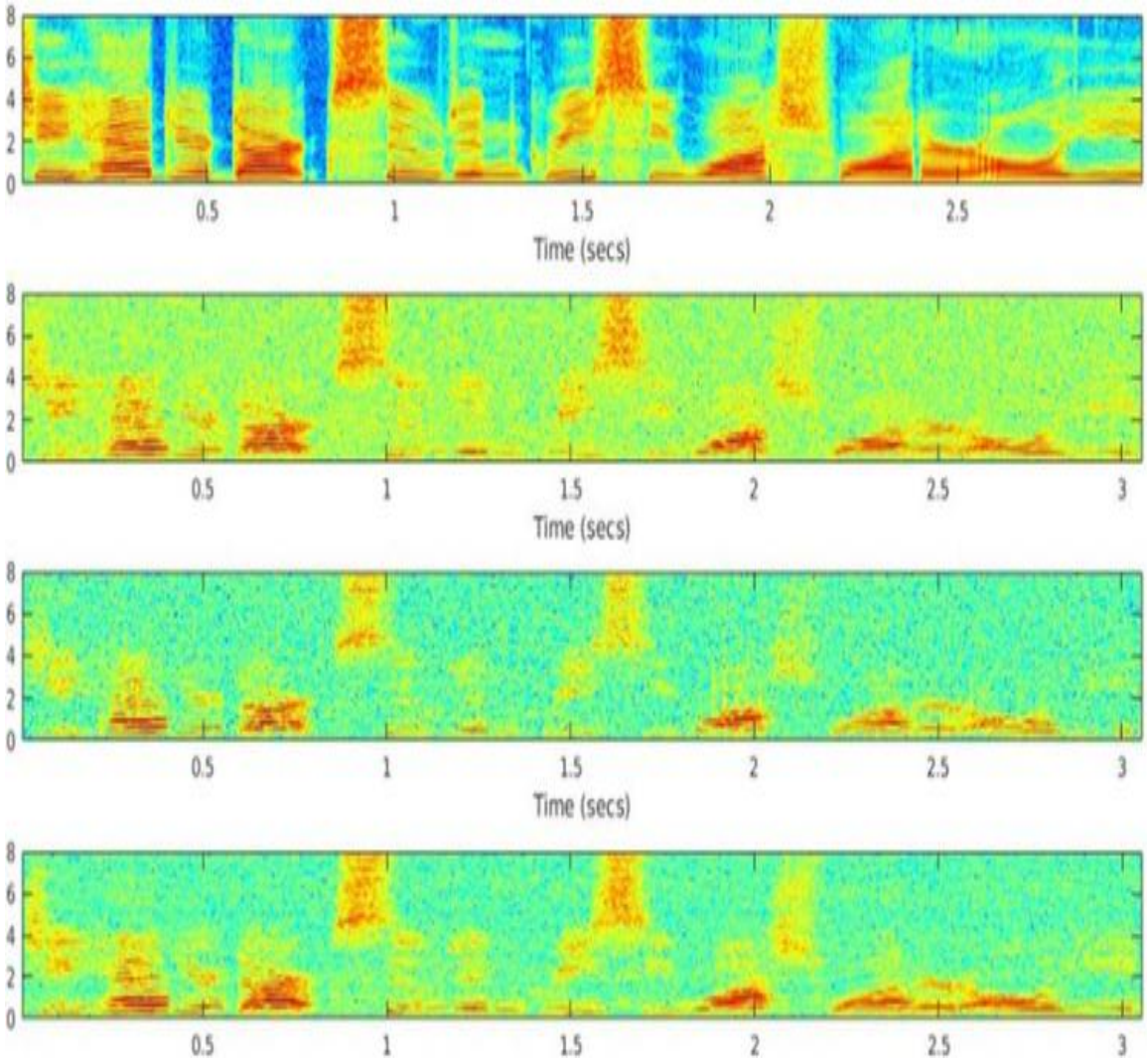

Figure 2: The spectrograms of (a) Clean speech signal (b) Microphone recording (c) Signal enhanced using conventional beamforming: Estimating DOA using the GCC-PHAT algorithm [11] followed by MVDR beamforming (d) Signal enhanced using [9] technique

MVDR beamforming was performed to enhance the signal after the CGMM steering vector estimation. In figure 2, we observe that CGMM outperforms the conventional DOA based techniques. Although both the algorithms attempt to denoise (evident from the blue patches), there is a lot of information loss in case of conventional algorithm. Observe that the high frequency signal component is retained to a much larger extent in CGMM approach. Even the signal characteristics at lower frequencies are preserved much better when compared to the GCC-PHAT DOA based method.

Signal	STOI Score	OPS Score
Microphone Recording	0.6226	22.8711
`DOA_MVDR` Output	0.6784	24.2194
`CGMM` Output	0.7064	29.5826

Table 2: Short Time Objective Intelligiblity (STOI) [19] and Overall Perceptual Score (OPS) [20] for original recordings, and signals enhanced using conventional and CGMM techniques

6 Future Work

Extending the architecture to the entire frequency range and further, speech enhancement using MVDR beamforming is to be incorporated. Considering the performance of the CGMM approach, and the effectiveness of our algorithm over it (for a smaller architecture), we believe that the proposed network would yield promising results on extension.

7 Acknowledgement

I would like to thank Prof. Rajesh M. Hegde for providing me this opportunity. I would also like to extend my gratitude to Vishnuvardhan Varanasi for his guidance at each and every step of the project.

References

[1] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120, 1979.

[2] Michael S Brandstein and Harvey F Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 1, pages 375–378. IEEE, 1997.

[3] Soumitro Chakrabarty, Emanuël Habets, et al. Broadband doa estimation using convolutional neural networks trained with noise signals. *arXiv preprint arXiv:1705.00919*, 2017.

[4] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 32(6):1109–1121, 1984.

[5] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 708–712. IEEE, 2015.

[6] Hakan Erdogan, John R Hershey, Shinji Watanabe, Michael I Mandel, and Jonathan Le Roux. Improved mvdr beamforming using single-channel mask prediction networks. In *INTERSPEECH*, pages 1981–1985, 2016.

[7] Szu-Wei Fu, Yu Tsao, and Xugang Lu. Snr-aware convolutional neural network modeling for speech enhancement. In *Interspeech*, pages 3768–3772, 2016.

[8] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach. Neural network based spectral mask estimation for acoustic beamforming. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 196–200. IEEE, 2016.

[9] Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka, and Tomohiro Nakatani. Robust mvdr beamforming using time-frequency masks for online/offline asr in noise. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5210–5214. IEEE, 2016.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[12] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777, 2014.

[13] Junfeng Li, Shuichi Sakamoto, Satoshi Hongo, Masato Akagi, and Yôiti Suzuki. Two-stage binaural speech enhancement with wiener filter for high-quality speech communication. *Speech Communication*, 53(5):677–689, 2011.

[14] Jae S Lim and Alan V Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

[15] Pascal Scalart et al. Speech enhancement based on a priori signal to noise estimation. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 629–632. IEEE, 1996.

[16] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986.

[17] Mehrez Souden, Shoko Araki, Keisuke Kinoshita, Tomohiro Nakatani, and Hiroshi Sawada. A multichannel mmse-based framework for speech source separation and noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1913–1928, 2013.

[18] Petre Stoica and Kenneth C Sharman. Maximum likelihood methods for direction-of-arrival estimation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(7):1132–1143, 1990.

[19] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

[20] Emmanuel Vincent. Improved perceptual metrics for the evaluation of audio source separation. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 430–437. Springer, 2012.

[21] Dang Hai Tran Vu and Reinhold Haeb-Umbach. Blind speech separation employing directional statistics in an expectation maximization framework. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 241–244. IEEE, 2010.

[22] DeLiang Wang. Deep learning reinvents the hearing aid. *IEEE Spectrum*, 54(3):32–37, 2017.

[23] Kevin W Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4029–4032. IEEE, 2008.