# DNN-guided Parameter Estimation for Speech Enhancement

October 15, 2024

Shuai Tao[1], Pejman Mowlaee[2], Jesper Rindom Jensen[1], Mads Græsbøll Christensen[1]

[1] Audio Analysis Lab, Electronic Systems, Aalborg University, Denmark
[2]GN Audio A/S, Lautrupbjerg 7, 2750 Ballerup, Denmark

AALBORG UNIVERSITY
DENMARK

# Agenda

# Outline

# Introduction



- Minimum variance distortionless response (MVDR) provides an optimal solution for beamforming.
- Statistics i.e. noise power density (PSD) matrix and steering vector are required for MVDR beamforming.
- Knowledge of the SPP allows for better estimation of noise and speech statistics. For multi-channel, SPP can be estimated as more effective as it includes spatial information.
- To achieve accurate MC-SPP estimation, the low-parameter model is employed to estimate MC-SPP.
- Guided by the MC-SPP estimate, statistics are estimated. Finally, an improved MVDR beamforming is performed.

# Outline

# Signal Model and MC-SPP



In the short-time Fourier transform domain, with noise and reverberation, a microphone array ($M$ microphones) observed one signal $\mathbf{y}(k,l) = [y_0(k,l), .., y_{M-1}(k,l)]^T$, which is given by

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{v}(k,l) + \mathbf{r}(k,l), \tag{1}$$

where $k$ and $l$ are the frequency and time index, $\mathbf{x}(k,l) = [x_0(k,l), ..., x_{M-1}(k,l)]^T$ is the clean speech, $\mathbf{v}(k,l) = [v_0(k,l), ..., v_{M-1}(k,l)]^T$ is the background noise, and $\mathbf{r}(k,l) = [r_0(k,l), ..., r_{M-1}(k,l)]^T$ is the reverberant speech. Assuming $\mathbf{n}(k,l) = \mathbf{r}(k,l) + \mathbf{v}(k,l)$, (1) is rewritten as

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{n}(k,l). \tag{2}$$

# Signal Model and MC-SPP



The clean and noise power spectral density (PSD) matrices are given by

$$\Phi_{xx}(k,l) = \mathrm{E}\left[\mathbf{x}(k,l)\mathbf{x}^H(k,l)\right], \tag{3}$$

and

$$\Phi_{nn}(k,l) = \mathrm{E}\left[\mathbf{n}(k,l)\mathbf{n}^H(k,l)\right]. \tag{4}$$

Let $\mathcal{H}_0$ and $\mathcal{H}_1$ denote the speech absence and presence, we have

$$\begin{cases} \mathcal{H}_0 : \mathbf{y}(k,l) = \mathbf{n}(k,l) \\ \mathcal{H}_1 : \mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{n}(k,l). \end{cases} \tag{5}$$

# Signal Mode and MC-SPP



With the multivariate Gaussian distribution [1], the likelihood functions of speech and noise can be obtained by

$$p[\mathbf{y}(k,l)|\mathcal{H}_1] = \frac{1}{\pi^M \det\left[\Phi_{xx}(k,l) + \Phi_{nn}(k,l)\right]} \times \exp\{-\mathbf{y}(k,l)\left[\Phi_{nn}(k,l) + \Phi_{xx}(k,l)\right]^{-1}\mathbf{y}(k,l)\}, \tag{6}$$

and

$$p[\mathbf{y}(k,l)|\mathcal{H}_0] = \frac{1}{\pi^M \det\left[\Phi_{nn}(k,l)\right]} \times \exp\{-\mathbf{y}^H(k,l)\Phi_{nn}(k,l)\mathbf{y}(k,l)\}. \tag{7}$$

# Signal Mode and MC-SPP



Using Bayes rule, the *a posteriori* MC-SPP $p(k,l) = p[\mathbf{y}(k,l)|\mathcal{H}_1]$ can be obtained by

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}[1+\xi(k,l)]\exp\left[-\frac{\beta(\mathrm{k,l})}{1+\xi(\mathrm{k,l})}\right]\right\}^{-1}, \quad (8)$$

where $q(k,l)$ is the *a priori* speech absence probability (SAP), $\xi(k,l)$ is the multi-channel *a priori* signal-to-noise ratio (SNR) defined as

$$\xi(k,l) = \mathrm{tr}\left[\Phi_{nn}^{-1}(k,l)\Phi_{xx}(k,l)\right], \quad (9)$$

and $\beta(k,l)$ is given by

$$\beta(k,l) = \mathbf{y}^H(k,l)\Phi_{nn}^{-1}(k,l)\Phi_{xx}(k,l)\Phi_{nn}^{-1}(k,l)\mathbf{y}(k,l). \quad (10)$$

Submitting (9) and (10) to (8), the *a posteriori* MC-SPP is obtained.

# Outline

# Proposed Method

## Learning-based MC-SPP estimation



One low-parameter DNN model $F$ with parameter $\Theta$ is employed to estimate MC-SPP, we have

$$\hat{p}(k,l) = F^{\Theta}\{\mathbf{y}(k,l)\}, \tag{11}$$

where $\hat{p}(k,l)$ is the MC-SPP estimate.
During training, the *a priori* SAP is computed directly from the training data. It is obtained as

$$q(k,l) = 1 - \frac{\mathrm{tr}\left[\Phi_{xx}(k,l)\right]}{\mathrm{tr}\left[\Phi_{xx}(k,l)\right] + \mathrm{tr}\left[\Phi_{nn}(k,l)\right]}. \tag{12}$$

Additionally, the loss function is Kullback-Leibler divergence, which is given by

$$\mathcal{L}(p(k,l), \hat{p}(k,l)) = p(k,l)\log\left(\frac{p(k,l)}{\hat{p}(k,l)}\right). \tag{13}$$

# Proposed Method
## Statistics estimation



With the MC-SPP estimate in (11), the noise and clean PSD matrices are updated recursively,

$$\hat{\Phi}_{nn}(k,l) = \alpha_n \Phi_{nn}(k,l-1) + (1-\alpha_n)(1-\hat{p}(k,l))\mathbf{y}(k,l)\mathbf{y}^H(k,l), \quad (14)$$

$$\hat{\Phi}_{xx}(k,l) = \alpha_x \Phi_{xx}(k,l-1) + (1-\alpha_x)\hat{p}(k,l)\mathbf{y}(k,l)\mathbf{y}^H(k,l), \quad (15)$$

where $0 < \alpha_N < 1$ and $0 < \alpha_x < 1$. With the $M$ dimensional selection vector $\mathbf{u}_1 = [1, 0, ..., 0]^T$, the steering vector is given by

$$\mathbf{d}(k,l) = \frac{\hat{\Phi}_{xx}(k,l)\mathbf{u}_1}{\mathbf{u}_1^H \hat{\Phi}_{xx}(k,l)\mathbf{u}_1} \quad (16)$$

# Proposed Method
## Improved MVDR Beamforming



Then the minimum variance distortionless response (MVDR) weight is computed by

$$\mathbf{h}_{\text{MVDR}} = \frac{\Phi_{nn}^{-1}(k,l)\mathbf{d}(k,l)}{\mathbf{d}^{H}(k,l)\Phi_{nn}^{-1}(k,l)\mathbf{d}(k,l)}, \tag{17}$$

. To improve MVDR performance, with MC-SPP estimate, one modification is given by

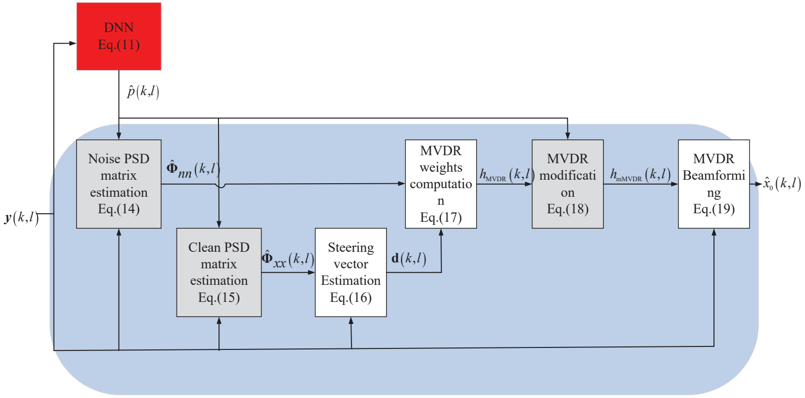
$$\mathbf{h}_{\text{mMVDR}}(k,l) = \hat{p}(k,l)\mathbf{h}_{\text{MVDR}}^{H}(k,l). \tag{18}$$

With (18), beamforming can be performed by

$$\hat{x}_0(k,l) = \mathbf{h}_{\text{mMVDR}}^{H}(k,l)\mathbf{y}(k,l). \tag{19}$$

Finally, submitting (18) to (19), the desired speech of the first channel can be obtained.

# Proposed Method
## DNN-guided MVDR Beamforming




MC-SPP-based MVDR beamforming

Figure 1: The pipeline of the proposed learning-based *a posteriori* MC-SPP estimation for multi-channel speech enhancement. The DNN-based MC-SPP estimate guides the estimation of the statistics in the grey box.

# Outline

# Experimental Settings



- Scalar temporal averaging factors in (14) and (15): $\alpha_n = 0.99$ and $\alpha_x = 0.1$.
- Noise type: diffuse noise is generated in the isotropic environment [2].

Table 1: Acoustic parameter configurations for experiment data generation

| | |
|---|---|
| Speech dataset | DNS Challenge 2021, read speech |
| Noise dataset | Audioset Freesound, Demand |
| Room size | Length: 10m; width: 8m; height: 3m |
| Microphone Array | Linear array with 6 microphones |
| 2*Array position | First microphone: [5, 1.5, 1.7]<br>10 cm distance with others |
| Source position | from [5, 2.75, 1.7] to [8, 4.75, 1.7] |
| $RT_{60}$ | 0.3s |

# Outline

# Experimental Results
## Spectrogram



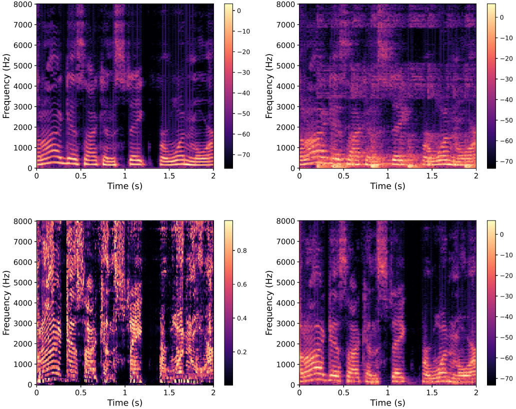



Figure 2: From left to right, in the first row, there are clean and noisy speech. In the second row, the MC-SPP estimate and enhanced speech.

# Experimental Results
## Numerical Results



Table 2: Multi-channel speech enhancement performance comparison.

| | PESQ | STOI | DNSMOS | Param | MB/s |
|---|---|---|---|---|---|
| Noisy | 1.42 | 0.632 | 2.49 | - | - |
| M1 [2] | 1.46 | 0.673 | 2.44 | - | - |
| DNN-CRNN | | | | | |
| M2 [3] | 1.57 | 0.644 | 2.38 | 4.59 M | 24 |
| Ours | 2.04 | 0.777 | 2.58 | 4.59M | 24 |
| DNN-Conformer | | | | | |
| M3 [4] | 1.58 | 0.732 | 2.40 | 8.44 M | 69 |
| Ours (18) | 1.82 | 0.723 | 2.54 | 3.79M | 52 |
| DNN-DeFT-AN | | | | | |
| M4 [5] | **2.21** | 0.783 | 2.69 | 0.68 M | 1007 |
| Ours + [6] | 1.44 | 0.707 | 2.31 | 0.68M | 994 |
| Ours (18) | **2.21** | **0.803** | **2.72** | 0.68 M | 994 |

# Outline

# Conclusion and Discussion



- Achieved accurate MC-SPP estimation using the DNN.
- More accurate MC-SPP estimate can guide more accurate statistics estimation, i.e., noise PSD matrix and steering vector.
- Our proposed method outperforms other benchmarks in terms of enhanced speech quality.
- For our proposed method, we give a better solution that can achieve higher performance with the same DNN model.

# References



[1] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, 2009.

[2] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[3] S. Chakrabarty and E. A. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[4] M. Kim, S. Cheong, and J. W. Shin, "DNN-based Parameter Estimation for MVDR Beamforming and Post-filtering," in *Proc. INTERSPEECH 2023*, 2023, pp. 3879–3883.

[5] D. Lee and J.-W. Choi, "DeFT-AN: Dense frequency-time attentive network for multichannel speech enhancement," *IEEE Signal Processing Letters*, vol. 30, pp. 155–159, 2023.

[6] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, 2011.

Thank you!

AALBORG UNIVERSITY
DENMARK