# ON SDW-MWF AND VARIABLE SPAN LINEAR FILTER WITH APPLICATION TO SPEECH RECOGNITION IN NOISY ENVIRONMENTS

*Ziteng Wang, Lu Yin, Junfeng Li and Yonghong Yan*

Key Laboratory of Speech Acoustics and Content Understanding,
Institute of Acoustics, University of Chinese Academy of Sciences, Beijing, China

## ABSTRACT

Neural network based spectral mask estimation for acoustic beamforming, which consists of linear filtering and mask estimation, has shown to be a promising approach for robust speech recognition in noisy environments. Nevertheless, few improvements are made on the linear filtering. In this paper, we investigate the Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) and the variable span linear filter, and prove that they can be linked by Generalized Eigenvalue Decomposition (GEVD) of the speech covariance matrix. The resulting GEVD based SDW-MWF largely reduces the word error rate and even achieves competitive recognition performance with the state-of-the-art generalized eigenvalue beamformer. Furthermore, we found that the recent signal approximation is no better than mask approximation when combined in calculating the linear filter coefficients.

***Index Terms***— multichannel Wiener filter, variable span filter, neural network based masking, speech recognition

## 1. INTRODUCTION

The integration of multichannel linear filters and Deep Neural Network (DNN) based spectral mask estimation has shown to be a promising approach for robust automatic speech recognition in the CHiME challenges [1, 2, 3]. Many recent advances originated from it, such as beamnet[4], multichannel end-to-end training [5] and deep clustering based beamforming [6, 7]. This approach can be analyzed from two perspectives: linear filtering and mask estimation. Much work has focused on the mask prediction while less attention was paid to the linear filtering.

The Generalized Eigenvalue (GEV) beamformer [8] and Minimum Variance Distortionless Response (MVDR) [9] become popular, because they are independent of the microphone array geometry and just rely on the speech and noise second-order statistics. Nevertheless, these two beamformers are optimized under different constraints. GEV is designed to achieve maximum output Signal-to-Noise Ratio (SNR) while MVDR is to achieve minimum speech distortion. Meanwhile, the Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF), that is the optimal solution under the weighted Minimum Mean Squared Error (MMSE) criterion, has also attracted many interest in the past decades [10, 11, 12, 13]. In [14], these filters are compared in terms of the relative speech recognition performances together with a recently proposed Variable Span (VS) linear filter [15]. The SDW-MWF is found to underperform the others.

Recent work has concluded that the linear filters are all equivalent up to a scaling factor if they are formulated under the same framework [16, 17]. For instance, under the narrowband approximation, SDW-MWF can be reformulated into the form of MVDR or plain MWF by setting its trade-off parameter to 0 and 1, respectively [11]. Various filter variants are also derived for the VS filter [15]. In this paper, we will further show that the SDW-MWF is linked to the VS filter using the Generalized Eigenvalue Decomposition (GEVD) of the speech covariance matrix. The GEVD based SDW-MWF is able to achieve competitive speech recognition performance with the others.

When calculating the beamformer coefficients, the speech and noise covariance matrices need to be estimated from the noisy observations. This is where the emerging mask prediction techniques step in, since the masks can be interpreted as time-frequency signal presence probabilities. The idea of masking originates from the early work of Computational Auditory Scene Analysis (CASA) [18]. With deep learning, mask prediction is treated as a classification task and different training features and targets are thoroughly studied [19, 20]. Recent advances, such as deep clustering [21] and permutation invariant training [22], address the task from the segregation and separation perspectives. Therefore, it is quite intuitive to integrate these new techniques with linear filtering as what has been done in [6]. Note that these methods are single channel based, one question remains that whether the improvements on the mask prediction can bring according improvements to the linear filters. Specially, it is concluded that Signal Approximation (SA) is a better objective than Mask Approximation (MA) in source separation tasks [23]. We are motivated to investigate the SA based

This work is partially supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Nos. XDA06030100, XDA06030500, XDA06040603) National 863 Program (No. 2015AA016306) and National 973 Program (No. 2013CB329302).

linear filters and compare them with the MA based ones in the recent CHiME-4 speech recognition task.

## 2. SIGNAL MODEL

In a typical enclosure, a speech signal $s$ is emitted from the source and captured by a microphone array with $M$ microphones. The observations at time $t$ are written as

$$y_m(t) = g_m(t) \odot s(t) + n_m(t), \quad m = 1, 2, ..., M \tag{1}$$

where $\odot$ denotes convolution, $g_m$ is the acoustic impulse response relating the source and the $m$th microphone and $n_m$ is the additive noise, which is assumed to be uncorrelated with speech. In the Short-time Fourier (STFT) domain, the microphone signals are given by

$$Y_m(l,k) = X_m(l,k) + N_m(l,k) \tag{2}$$

where $X_m(l,k) = G_m(k)S(l,k)$ is the source image in the $m$th microphone under the narrowband approximation. $l$ is the frame index and $k$ is the frequency index. Note that the subsequent operations are performed on time-frequency bin basis, thus the indexes will be omitted for clarity.

## 3. ACOUSTIC BEAMFORMERS

Beamforming techniques aim to design a complex-valued filter $\mathbf{h} = [H_1,\ H_2, ...,\ H_M]^T$ that extracts the desired source and suppresses the other interfering components. $^T$ denotes transposition. The filter is applied to the observation vector and the output is

$$\begin{aligned} O &= \mathbf{h}^H \mathbf{y} \\ &= \mathbf{h}^H \mathbf{x} + \mathbf{h}^H \mathbf{n} \end{aligned} \tag{3}$$

where $^H$ denotes Hermitian transpose, $\mathbf{y} = [Y_1,\ Y_2, ...,\ Y_M]^T$, $\mathbf{x} = [X_1,\ X_2, ...,\ X_M]^T$ and $\mathbf{n} = [N_1,\ N_2, ...,\ N_M]^T$.

### 3.1. SDW-MWF [10]

The SDW-MWF is derived under the weighted MSE criterion with respect to an arbitrary channel of the reverberated source, say $X_1$:

$$\min_{\mathbf{h}} E\{|\mathbf{h}^H \mathbf{x} - X_1|^2\} + \mu E\{|\mathbf{h}^H \mathbf{n}|^2\} \tag{4}$$

where $\mu \geq 0$ is known as the trade-off parameter that tunes speech distortion versus noise reduction. A larger $\mu$ will lead to more noise reduction at the expense of more speech distortion. The solution to (4) is

$$\mathbf{h}_{\text{SDW-MWF}} = (\mathbf{\Phi}_{xx} + \mu \mathbf{\Phi}_{nn})^{-1} \mathbf{\Phi}_{xx} \mathbf{u}_1 \tag{5}$$

where $\mathbf{\Phi}_{xx} = E\{\mathbf{x}\mathbf{x}^H\}$ is the speech covariance matrix, $\mathbf{\Phi}_{nn} = E\{\mathbf{n}\mathbf{n}^H\}$ is the noise covariance matrix and $\mathbf{u}_1 = [1,\ 0,\ ...,\ 0]^T$ is an $M$-dimensional vector that projects on the first channel.

### 3.2. Variable span filter [15]

Assuming $\mathbf{\Phi}_{nn}$ is of full rank and $\mathbf{\Phi}_{xx}$ is of rank-$P$, the two Hermitian matrices can be jointly diagonalized as follows:

$$\begin{cases} \mathbf{B}^H \mathbf{\Phi}_{xx} \mathbf{B} = \mathbf{\Lambda} \\ \mathbf{B}^H \mathbf{\Phi}_{nn} \mathbf{B} = \mathbf{I} \end{cases} \tag{6}$$

where $\mathbf{B}$, invertible but not necessarily orthogonal, is the eigenvector matrix of $\mathbf{\Phi}_{nn}^{-1}\mathbf{\Phi}_{xx}$. $\mathbf{I}$ is the $M \times M$ identity matrix. $\Lambda$ is a diagonal matrix whose elements are the eigenvalues arranged in descending order $\lambda_1 \geq \lambda_2 \geq \cdots > \lambda_{P+1} = \cdots = \lambda_M = 0$. That is to say, the last $M - P$ eigenvalues are exactly zero while its first $P$ eigenvalues are positive. The corresponding eigenvectors are denoted by $\mathbf{b}_1, \mathbf{b}_2, ...\mathbf{b}_M$, which form a new basis in the signal space. Then it is always possible to write $\mathbf{h}$ as

$$\mathbf{h} = \mathbf{B}\mathbf{a} \tag{7}$$

where $\mathbf{a}$ are the coordinates in the new basis.

The VS trade-off filter is obtained by $\mathbf{a}_Q = (\Lambda_Q + \mu I_Q)^{-1} \mathbf{B}_Q^H \mathbf{\Phi}_{xx} \mathbf{u}_1$ with the last $M - Q$ components of $\mathbf{a}$ being 0. Accordingly,

$$\mathbf{h}_{\text{VS}} = \sum_{q=1}^{Q} \frac{\mathbf{b}_q \mathbf{b}_q^H}{\mu + \lambda_q} \mathbf{\Phi}_{xx} \mathbf{u}_1 \tag{8}$$

where $1 \leq Q \leq M$ denotes the span in the signal space.

### 3.3. Link between SDW-MWF and VS

It is only known that for $Q = M$, the VS trade-off filter is the same as the SDW-MWF since both are the optimal solutions under the weighted MSE constraint [15]. Here, we derive a more general link between the two filters by analyzing $\mathbf{\Phi}_{xx}$. Using a basis $\mathbf{p}_{1,2...,M}$, the speech covariance matrix can be decomposed as

$$\mathbf{\Phi}_{xx} = \underbrace{\sum_{q=1}^{Q} \sigma_{x_q} \mathbf{p}_q \mathbf{p}_q^H}_{\mathbf{\Phi}_Q} + \mathbf{\Phi}_Z \tag{9}$$

where $\sigma_{x_q}$ is the scaling factor, $\mathbf{\Phi}_Q$ is the rank-$Q$ approximation of the speech covariance matrix and $\mathbf{\Phi}_Z$ is a remainder matrix. From (6), we have

$$\mathbf{\Phi}_{xx} = \mathbf{B}^{-H} \Lambda \mathbf{B}^{-1} \tag{10}$$

$$\mathbf{\Phi}_Q = \mathbf{B}^{-H} \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_Q, 0, \ldots, 0\} \mathbf{B}^{-1} \tag{11}$$

$$\mathbf{\Phi}_Z = \mathbf{B}^{-H} \text{diag}\{0, \ldots, 0, \lambda_{Q+1}, \ldots, \lambda_M\} \mathbf{B}^{-1} \tag{12}$$

where diag$\{\}$ means diagonal matrix. The SDW-MWF formula (5) then leads to

$$\mathbf{h} = \mathbf{B} \left( \text{diag}\{\frac{1}{\mu + \lambda_q}\} \right) \mathbf{B}^{-1} \mathbf{u}_1 \tag{13}$$

Using $\mathbf{B}^{-1} = \mathbf{B}^H \mathbf{\Phi}_{xx}$, the above formulation is indeed the VS trade-off filter (8).

It is further noted that

$$\mathbf{\Phi}_{xx}\mathbf{u}_1 = \mathbf{\Phi}_Q \mathbf{u}_1 + \mathbf{\Phi}_Z \mathbf{u}_1 = \mathbf{\Phi}_Q \mathbf{u}_1 \tag{14}$$

Then (13) can be written as

$$\mathbf{h}_{\text{GEVD-SDW-MWF}} = (\mathbf{\Phi}_Q + \mu \mathbf{\Phi}_{nn})^{-1} \mathbf{\Phi}_Q \mathbf{u}_1 \tag{15}$$

By comparing (15) to (5), it is seen that the VS filter is equivalent to replacing $\mathbf{\Phi}_{xx}$ in SDW-MWF by the GEVD based $\mathbf{\Phi}_Q$, and the span is decided by the number of eigenvectors used in building the filter.

## 4. MASK ESTIMATION

The linear filters are specified as functions of $\mathbf{\Phi}_{xx}$ and $\mathbf{\Phi}_{nn}$, for which the estimation is based on neural networks. The procedure is illustrated in Fig. 1. A Bidirectional Long Short-term Memory (BLSTM) network is first trained on ideal binary mask targets, defined as

$$\mathcal{M}_x = \begin{cases} 1 & \text{SNR} > LC_x \\ 0 & \text{otherwise} \end{cases}, \; \mathcal{M}_n = \begin{cases} 1 & \text{SNR} > LC_n \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

where $LC_x$ and $LC_n$ are the speech and noise local thresholds, respectively. The training loss function is given by

$$\text{loss}_{\text{MA}} = \text{CE}(\widetilde{\mathcal{M}_x}, \mathcal{M}_x) + \text{CE}(\widetilde{\mathcal{M}_n}, \mathcal{M}_n) \tag{17}$$

where CE means the cross-entropy loss and tilde variables are predictions. This is referred as Mask Approximation (MA).

In [23], Signal Approximation (SA) was proposed and the training loss function was changed to

$$\text{loss}_{\text{SA}} = (\widetilde{\mathcal{M}_x} \cdot |Y| - |X|)^2 + (\widetilde{\mathcal{M}_n} \cdot |Y| - |N|)^2 \tag{18}$$

where $|X|$ is the reference speech magnitude spectrum and $|N|$ is the noise magnitude spectrum. SA was found better than MA because the optimizing objective (18) was more related to the speech separation task at hand. The SA setup also avoids the explicit design of the ideal masks. We thus first combine BLSTM and SA (BLSTM-SA) and followed by further combining with linear filters. The investigation is whether it can bring performance benefits to our speech recognition task. The Phase Sensitive Approximation (PSA) is evaluated together, which is given by replacing $|X|$ by $|X|cos(\theta_Y - \theta_X)$ and $|N|$ by $|N|cos(\theta_Y - \theta_N)$, with $\theta$ denoting the signal phase.

The masks are used to obtain $\widetilde{\mathbf{\Phi}_{xx}} = \frac{1}{\sum_l \widetilde{\mathcal{M}_x}} \sum_l \widetilde{\mathcal{M}_x} \mathbf{y}\mathbf{y}^H$ and $\widetilde{\mathbf{\Phi}_{nn}} = \frac{1}{\sum_l \widetilde{\mathcal{M}_n}} \sum_l \widetilde{\mathcal{M}_n} \mathbf{y}\mathbf{y}^H$.

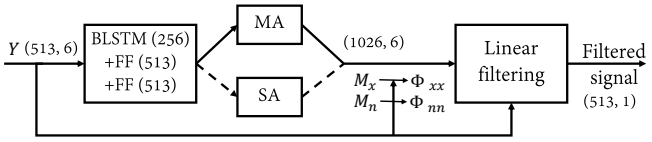


**Fig. 1**. Illustration of linear filtering supported by mask prediction neural network. The numbers in brackets denote the dimension of the variables or components.

## 5. SPEECH RECOGNITION EXPERIMENTS

### 5.1. Setup

The CHiME-4 dataset is used in the experiments. It contains both real and simulated data. The real data is recorded in four everyday environments: bus (BUS), cafe (CAF), pedestrian area (PED) and street junction (STR). The signals are recorded by six microphones and sampled at 16 kHz. The simulated data is generated by artificially mixing clean speech data from the WSJ0 corpus with noisy backgrounds. We use the baseline speech recognition system built with Kaldi (available at `https://github.com/kaldi-asr/kaldi/tree/master/egs/chime4`). The acoustic model is trained on the noisy training set while the linear filtering methods are applied only to the development set and test set. The filtered signals are then sent for transcription and the Word Error Rates (WERs) are computed.

For the front-end, the BLSTM-MA is adopted from [2], and the BLSTM-SA is extended on this. The network has one BLSTM layer with 256 nodes, and two feed forward (FF) layers each with 513 nodes. STFT is performed in 1024 points. The input is single channel magnitude spectrum of dimension 513 and the output is the concatenation of speech and noise masks. The network is initialized from Gaussian distributed samples and the Adam method is used for fine tuning. Sequence normalization and dropout are used as in the original setup. Early stopping is employed with a patience of 5 epochs.

The GEV filter is chosen as the baseline method, that is defined as

$$\mathbf{h}_{\text{GEV}} = \underset{\mathbf{h}}{\text{argmax}} \, \frac{\mathbf{h}^H \mathbf{\Phi}_{xx} \mathbf{h}}{\mathbf{h}^H \mathbf{\Phi}_{nn} \mathbf{h}} \tag{19}$$

for which we have $\mathbf{h}_{\text{GEV}} = \mathbf{b}_1$. Blind Analytic Normalization (BAN) can further be applied as a post filter to reduce speech distortion. The span for VS and GEVD-SDW-MWF is set to be $Q = 1$ with $\mu = 1$, which gives the best results. The rank-1 constraint also benefits other linear filters as shown latter.

### 5.2. WER results

The WER results obtained with the official CHiME-4 back-end speech recognizer are summarized in Table I. We first compare the results of BLSTM-SA/PSA with BLSTM-MA

**Table 1**. WERs of different filters obtained with the official CHiME-4 back-end speech recognizer. The right part gives details in different environments for the real test set. Bold indicates the best result.

| | Dataset | dev | | test | | test-real | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | simu | real | simu | real | BUS | CAF | PED | STR |
| BLSTM-SA | GEV-BAN | 4.83 | 4.53 | 6.68 | 8.10 | 11.61 | 6.78 | 7.32 | 6.69 |
| BLSTM-PSA | GEV-BAN | 5.14 | 4.83 | 6.59 | 7.67 | 10.58 | 6.31 | 7.42 | 6.35 |
| BLSTM-MA | GEV-BAN | 4.93 | 4.83 | 6.29 | 7.25 | 10.00 | 6.39 | 6.22 | 6.39 |
| | GEV | 5.25 | 4.76 | 6.98 | 7.14 | 8.34 | 6.72 | 7.04 | 6.46 |
| | VS | 3.83 | 4.45 | **4.15** | 7.35 | 11.46 | 5.96 | 5.85 | 6.13 |
| | SDW-MWF | 6.91 | 6.81 | 10.81 | 12.92 | 24.06 | 10.68 | 8.69 | 8.26 |
| | GEVD-SDW-MWF | 3.85 | 4.40 | 4.39 | 7.38 | 11.57 | 6.05 | 5.94 | 5.96 |
| BLSTM-MA | GEVD-GEV-BAN | **3.83** | **3.84** | 4.21 | 6.17 | 9.44 | **4.99** | **4.93** | 5.30 |
| | GEVD-GEV | 4.11 | 3.91 | 4.65 | **5.93** | **7.80** | 5.57 | 5.29 | **5.09** |
| | GEVD-VS | 3.88 | 4.07 | 4.42 | 6.86 | 11.07 | 5.66 | 5.42 | 5.30 |

using the GEV-BAN method. Though the SA/PSA objectives avoid the design of ideal masks and lead to better separation performance, the combination with linear filters does not achieve corresponding benefits, as indicated by the higher errors on the test set. It is possible that the masks indeed function as time-frequency signal presence probabilities and the MA objective is more proper for the linear filters. In the following experiments, the BLSTM-MA setup is used.

GEV and GEV-BAN lead to similar results though they perform differently in terms of speech distortion. The main difference is observed on the BUS data, which contains strong low frequency noises. The GEV filter happens to feature a constant residual noise power and the low frequency energies are highly suppressed [14]. The original SDW-MWF performs significantly worse than others. However, with the GEVD decomposition and reconstruction of the speech covariance matrix, the WERs largely decrease and are now competitive to GEV and VS on both real and simulated data. Different results are obtained for GEVD-SDW-MWF and VS, which should be due to their different formulations and that the inversion operation in (15) is numerically instable.

### 5.3. Rank-1 prior assumption of $\mathbf{\Phi}_{xx}$

The GEVD of the speech covariance matrix links SDW-MWF to VS. Meanwhile, the rank-1 decomposition of $\mathbf{\Phi}_{xx}$ can be interpreted as a prior assumption under the narrowband approximation (2), which indicates that the GEVD reconstructed speech covariance matrix can be directly incorporated in all other linear filters. This turns out effective as shown in the last three rows of Table I. The relative WER reductions are 17% for GEV and 7% for VS on the real test data. The approach is most effective on the real STR environment and on the simulated data, where the rank-1 constraint better matches the low reverberant data scenarios. As far as we know, the GEVD-GEV is the most effective filter ever reported on the CHiME-4 real test data.

## 6. CONCLUSION

We looked into the neural network based mask estimation for acoustic beamforming method, and adopted it to the SDW-MWF and VS filter. The SDW-MWF was proved to be equivalent to the VS filter by replacing its speech covariance matrix with the GEVD reconstructed one. The resulting GEV-SDW-MWF significantly reduced the WER on the CHiME-4 speech recognition task. Furthermore, the rank-1 GEVD decomposition was combined in other filters and led to the effective GEVD-GEV filter. We also investigated the BLSTM-SA/PSA setup and concluded that it was no better than BLSTM-MA when the masks were employed in combination with linear filters. Our code is available at `https://github.com/ZitengWang/nn_mask`.

## 7. REFERENCES


[1] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

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

[3] Hakan Erdogan, Tomoki Hayashi, John R Hershey, Takaaki Hori, Chiori Hori, Wei-Ning Hsu, Suyoun Kim, Jonathan Le Roux, Zhong Meng, and Shinji Watanabe,

"Multi-channel speech recognition: LSTMs all the way through," in *CHiME-4 Workshop*, 2016.

[4] Jahn Heymann, Lukas Drude, Christoph Boeddeker, Patrick Hanebrink, and Reinhold Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel asr system," in *ICASSP*, 2017, pp. 5325–5329.

[5] Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, and John R. Hershey, "Multichannel end-to-end speech recognition," in *International Conference on Machine Learning*, 2017, pp. 2632–2641.

[6] Takuya Higuchi, Keisuke Kinoshita, Marc Delcroix, Kateina molkov, and Tomohiro Nakatani, "Deep clustering-based beamforming for separation with unknown number of sources," in *INTERSPEECH*, 2017, pp. 1183–1187.

[7] Lukas Drude and Reinhold Haeb-Umbach, "Tight integration of spatial and spectral features for bss with deep clustering embeddings," in *INTERSPEECH*, 2017, pp. 2650–2654.

[8] Ernst Warsitz and Reinhold Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[9] Henry Cox, Robert M Zeskind, and Mark Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[10] Ann Spriet, Marc Moonen, and Jan Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[11] Mehrez Souden, Jacob Benesty, and Sofiene Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.

[12] Bram Cornelis, Marc Moonen, and Jan Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, 2011.

[13] Romain Serizel, Marc Moonen, Bas Van Dijk, and Jan Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, 2014.

[14] Ziteng Wang, Emmanuel Vincent, Romain Serizel, and Yonghong Yan, "Rank-1 constrained multichannel wiener filter for speech recognition in noisy environments," *Computer Speech & Language*, vol. 49, pp. 37–51, 2018.

[15] Jesper Rindom Jensen, Jacob Benesty, and Mads Græsbøll Christensen, "Noise reduction with optimal variable span linear filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 631–644, 2016.

[16] Jacob Benesty, Mehrez Souden, and Jingdong Chen, "A perspective on multichannel noise reduction in the time domain," *Applied Acoustics*, vol. 74, no. 3, pp. 343–355, 2013.

[17] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[18] Deliang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines*, pp. 181–197, 2005.

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

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

[21] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 31–35.

[22] Dong Yu, Morten Kolbk, Zheng Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[23] Felix Weninger, John R. Hershey, Jonathan Le Roux, and Bjrn Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Signal and Information Processing*, 2015, pp. 577–581.