EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

REVIEW Open Access

A brief overview of speech enhancement with linear filtering

Jacob Benesty[1,2†], Mads Græsbøll Christensen[1†], Jesper Rindom Jensen[1*†] and Jingdong Chen[3†]

Abstract

In this paper, we provide an overview of some recently introduced principles and ideas for speech enhancement with linear filtering and explore how these are related and how they can be used in various applications. This is done in a general framework where the speech enhancement problem is stated as a signal vector estimation problem, i.e., with a filter matrix, where the estimate is obtained by means of a matrix-vector product of the filter matrix and the noisy signal vector. In this framework, minimum distortion, minimum variance distortionless response (MVDR), tradeoff, maximum signal-to-noise ratio (SNR), and Wiener filters are derived from the conventional speech enhancement approach and the recently introduced orthogonal decomposition approach. For each of the filters, we derive their properties in terms of output SNR and speech distortion. We then demonstrate how the ideas can be applied to single- and multichannel noise reduction in both the time and frequency domains as well as binaural noise reduction.

Keywords: Noise reduction; Speech enhancement; Orthogonal decomposition; Performance measures; Optimal linear filtering; Single-channel; Multichannel; Binaural; Time domain; Frequency domain

1 Review

1.1 Introduction

The problem of speech enhancement, or noise reduction as it is also sometimes called, is a well-known, longstanding problem with important applications in, for example, speech communication systems and hearing aids, where additive noise can, and often does, have a detrimental impact on the speech quality. Although the problem is a classical one and many solutions have been proposed throughout the years, it has arguably not been well-understood, even for the comparably simple case of linear filters. Indeed, it is not until quite recently that steps have been taken to accurately formulate the problem and characterize the desirable properties of possible solutions. Simply put, the performance of speech enhancement methods can be assessed in terms of two quantities, namely noise reduction and speech distortion, and an optimal solution to the speech enhancement would, thus, explicitly take both into account. As an example that this has historically not been done, consider the classical and well-known Wiener filter. It is usually derived from a mean-square error (MSE) criterion, and it is not until recently that its properties in terms of noise reduction and speech distortion have been thoroughly analyzed [1]. Although it was shown in [1] that the Wiener filter indeed improves the signal-to-noise ratio (SNR) (this had not been shown before), it does so without any explicit control over the amount of distortion incurred on the speech signal. As a result, other filters are worth considering. The minimum variance distortionless response (MVDR) filter, in principle, guarantees that no distortion is incurred on the speech signal while the noise is reduced as much as possible. The maximum SNR filter, on the other hand, yields the highest possible output SNR but does so at the cost of a considerable amount of speech distortion. Other competing methods to linear filtering include spectral subtraction methods [2], subspace methods [3,4], and statistical methods [5-7].

In this paper, we continue the research into methods for speech enhancement based on linear filtering. More specifically, we provide a brief overview of linear filters derived from the conventional approach and the recently introduced orthogonal decomposition approach. We do so in a more general framework than what is typical. More specifically, the speech enhancement problem

*Correspondence: jrj@create.aau.dk
†Equal contributors
[1]Audio Analysis Lab, AD:MT, Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark
Full list of author information is available at the end of the article

is stated as the problem of finding a rectangular filter matrix for estimating the speech signal vector from a noisy signal observation vector. Using the two aforementioned approaches, we derive the maximum SNR, Wiener, tradeoff, and MVDR filters and analyze and relate their properties. All of the derived filters are based on second-order statistics of the observed signal as well as the noise. While estimation of these statistics are not considered herein, there exist multiple, well-known methods for conducting this estimation in practice both in single-channel [8,9] and multichannel [10-12] scenarios. Finally, we then proceed to demonstrate and discuss their application in various settings, including time and frequency domain enhancement and single- and multichannel enhancement.

The rest of the paper is organized as follows. In Section 1.2, we introduce the signal model and basic assumptions and state the problem at hand. We then, in Section 1.3, address the problem using the conventional approach, define various useful performance measures, and derive and compare some optimal filters. We then proceed to present an alternative approach based on the orthogonal decomposition in Section 1.4, and we use this to derive optimal filters. These are then also compared in terms of their noise reduction and speech distortion properties. In Section 1.5, we show how the two approaches can be applied in various speech enhancement contexts before concluding on the work in Section 2.

1.2 Signal model and problem formulation

We consider the very general signal model of an observation signal vector of length M:

$$\begin{aligned}\mathbf{y} &= \left[y_1 \ y_2 \ \cdots \ y_M \right]^T \\ &= \mathbf{x} + \mathbf{v}, \end{aligned} \tag{1}$$

where the superscript ^{T} is the transpose operator, and $\mathbf{x}$ and $\mathbf{v}$ are the speech and noise signal vectors, respectively, which are defined similarly to the noisy signal vector, $\mathbf{y}$. We assume that the components of the two vectors $\mathbf{x}$ and $\mathbf{v}$ are zero mean, stationary, and circular. We further assume that these two vectors are uncorrelated, i.e., $E\left(\mathbf{x}\mathbf{v}^H\right) = E\left(\mathbf{v}\mathbf{x}^H\right) = \mathbf{0}_{M\times M}$, where $E(\cdot)$ denotes mathematical expectation, the superscript ^{H} is the conjugate-transpose operator, and $\mathbf{0}_{M\times M}$ is a matrix of size $M \times M$ with all its elements equal to 0. In this context, the correlation matrix (of size $M \times M$) of the observations is:

$$\begin{aligned}\Phi_{\mathbf{y}} &= E\left(\mathbf{y}\mathbf{y}^H\right) \\ &= \Phi_{\mathbf{x}} + \Phi_{\mathbf{v}}, \end{aligned} \tag{2}$$

where $\Phi_{\mathbf{x}} = E\left(\mathbf{x}\mathbf{x}^H\right)$ and $\Phi_{\mathbf{v}} = E\left(\mathbf{v}\mathbf{v}^H\right)$ are the correlation matrices of $\mathbf{x}$ and $\mathbf{v}$, respectively. In the rest of this paper, we assume that the rank of the speech correlation matrix, $\Phi_{\mathbf{x}}$, is equal to $P \leq M$ and the rank of the noise correlation matrix, $\Phi_{\mathbf{v}}$, is equal to M.

In order to be able to derive appropriate performance measures and optimal linear filters that can achieve a clear objective according to these measures, it is of great importance to define, without any ambiguity, the desired signal that we want to estimate or extract from the observations. Also, in general, $\mathbf{y}$ should be written explicitly as a function of this desired signal. In some context, x_1, the first element of $\mathbf{x}$, is the desired signal; in some other situations, the whole vector $\mathbf{x}$ or part of it is the desired signal vector. Therefore, in a general manner, our desired signal vector is defined as:

$$\mathbf{x}_Q = \left[x_1 \ x_2 \ \cdots \ x_Q \right]^T, \tag{3}$$

where $1 \leq Q \leq M$. In the same way, we define the vector $\mathbf{v}_Q$ as the first Q components of $\mathbf{v}$. Then, the objective of speech enhancement (or noise reduction) is to estimate $\mathbf{x}_Q$ from $\mathbf{y}$. This should be done in such a way that the noise is reduced as much as possible with no or little distortion of the desired signal vector [1,13-15]. In the rest of this study, we consider two important cases: without (conventional approach) and with the orthogonal decomposition of the speech signal vector.

1.3 Speech enhancement with the conventional approach

1.3.1 Principle

Our objective is to estimate $\mathbf{x}_Q$ from $\mathbf{y}$, even though $\mathbf{y}$ is not an explicit function of $\mathbf{x}_Q$. With linear filtering techniques [3,4,16-20], the desired signal vector is estimated as:

$$\begin{aligned}\mathbf{z} &= \mathbf{H}\mathbf{y} \\ &= \mathbf{H}\left(\mathbf{x} + \mathbf{v}\right) \\ &= \mathbf{x}_{\text{fd}} + \mathbf{v}_{\text{rn}}, \end{aligned} \tag{4}$$

where $\mathbf{z}$ is supposed to be the estimate of $\mathbf{x}_Q$,

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1^H \\ \mathbf{h}_2^H \\ \vdots \\ \mathbf{h}_Q^H \end{bmatrix} \tag{5}$$

is a rectangular filtering matrix of size $Q \times M$, $\mathbf{h}_q$, $q = 1, 2, \ldots, Q$ are complex-valued filters of length M, $\mathbf{x}_{\text{fd}} = \mathbf{H}\mathbf{x}$ is the filtered desired signal, and $\mathbf{v}_{\text{rn}} = \mathbf{H}\mathbf{v}$ is the residual noise. We deduce that the correlation matrix of $\mathbf{z}$ is:

$$\Phi_{\mathbf{z}} = \Phi_{\mathbf{x}_{\text{fd}}} + \Phi_{\mathbf{v}_{\text{rn}}}, \tag{6}$$

where $\Phi_{\mathbf{x}_{\text{fd}}} = \mathbf{H}\Phi_{\mathbf{x}}\mathbf{H}^H$ and $\Phi_{\mathbf{v}_{\text{rn}}} = \mathbf{H}\Phi_{\mathbf{v}}\mathbf{H}^H$.

An interesting particular case is $Q = P = 1$. In this scenario, Equation 4 simplifies to:

$$z = \mathbf{h}^H\mathbf{y}, \tag{7}$$

where $\mathbf{h}$ is a complex-valued filter of length M. Since $\Phi_{\mathbf{x}}$ is a rank 1 matrix, it can be written as:

$$\Phi_{\mathbf{x}} = \phi_{x_1}\mathbf{d}\mathbf{d}^H, \tag{8}$$

where $\phi_{x_1} = E\left(|x_1|^2\right)$ is the variance of x_1 and $\mathbf{d}$ is a vector of length M, whose first element is equal to 1.

1.3.2 Performance measures

We are now ready to define the most important performance measures in the context of linear filtering.

The input SNR is defined as:

$$\text{iSNR} = \frac{\text{tr}\left(\Phi_{\mathbf{x}_Q}\right)}{\text{tr}\left(\Phi_{\mathbf{v}_Q}\right)}, \tag{9}$$

where $\text{tr}(\cdot)$ denotes the trace of a square matrix, and $\Phi_{\mathbf{x}_Q}$ and $\Phi_{\mathbf{v}_Q}$ are the correlation matrices (of size $Q \times Q$) of $\mathbf{x}_Q$ and $\mathbf{v}_Q$, respectively.

The output SNR, obtained from Equation 6, helps quantify the SNR after filtering. It is given by:

$$\begin{aligned}\text{oSNR}\left(\mathbf{H}\right) &= \frac{\text{tr}\left(\Phi_{\mathbf{x}_{\text{fd}}}\right)}{\text{tr}\left(\Phi_{\mathbf{v}_{\text{rn}}}\right)} \\ &= \frac{\text{tr}\left(\mathbf{H}\Phi_{\mathbf{x}}\mathbf{H}^H\right)}{\text{tr}\left(\mathbf{H}\Phi_{\mathbf{v}}\mathbf{H}^H\right)}.\end{aligned} \tag{10}$$

Then, the main objective of speech enhancement is to find an appropriate $\mathbf{H}$ that makes the output SNR greater than the input SNR. Consequently, the quality of the noisy signal may be enhanced.

The noise reduction factor quantifies the amount of noise being rejected by $\mathbf{H}$. This quantity is defined as the ratio of the power of the original noise over the power of the noise remaining after filtering, i.e.,

$$\xi_{\text{nr}}\left(\mathbf{H}\right) = \frac{\text{tr}\left(\Phi_{\mathbf{v}_Q}\right)}{\text{tr}\left(\mathbf{H}\Phi_{\mathbf{v}}\mathbf{H}^H\right)}. \tag{11}$$

Any good choice of $\mathbf{H}$ should lead to $\xi_{\text{nr}}\left(\mathbf{H}\right) \geq 1$, in which case the noise has been attenuated.

The desired speech signal can be distorted by the rectangular filtering matrix. Therefore, the speech reduction factor is defined as:

$$\xi_{\text{sr}}\left(\mathbf{H}\right) = \frac{\text{tr}\left(\Phi_{\mathbf{x}_Q}\right)}{\text{tr}\left(\mathbf{H}\Phi_{\mathbf{x}}\mathbf{H}^H\right)}. \tag{12}$$

For optimal filters, we should have $\xi_{\text{sr}}\left(\mathbf{H}\right) \geq 1$ as the optimal filter would otherwise amplify the desired signal.

By making the appropriate substitutions, one can derive the relationship among the measures defined so far, i.e.,

$$\frac{\text{oSNR}\left(\mathbf{H}\right)}{\text{iSNR}} = \frac{\xi_{\text{nr}}\left(\mathbf{H}\right)}{\xi_{\text{sr}}\left(\mathbf{H}\right)}. \tag{13}$$

This fundamental expression indicates the equivalence between gain/loss in SNR and distortion (for both speech and noise).

Another way to measure the distortion of the desired signal vector due to the filtering operation is via the speech distortion index defined as:

$$\upsilon_{\text{sd}}\left(\mathbf{H}\right) = \frac{E\left[\left(\mathbf{x}_{\text{fd}} - \mathbf{x}_Q\right)^H\left(\mathbf{x}_{\text{fd}} - \mathbf{x}_Q\right)\right]}{\text{tr}\left(\Phi_{\mathbf{x}_Q}\right)}. \tag{14}$$

The speech distortion index is always greater than or equal to 0 and should be upper bounded by 1 for optimal rectangular filtering matrices[a]; so the higher the value of $\upsilon_{\text{sd}}\left(\mathbf{H}\right)$ is, the more the desired signal is distorted.

We define the error signal vector between the estimated and desired signals as:

$$\begin{aligned}\mathbf{e} &= \mathbf{z} - \mathbf{x}_Q \\ &= \mathbf{H}\mathbf{y} - \mathbf{x}_Q,\end{aligned} \tag{15}$$

which can also be expressed as the sum of two uncorrelated error signal vectors:

$$\mathbf{e} = \mathbf{e}_{\text{ds}} + \mathbf{e}_{\text{rs}}, \tag{16}$$

where

$$\mathbf{e}_{\text{ds}} = \mathbf{x}_{\text{fd}} - \mathbf{x}_Q \tag{17}$$

is the signal distortion due to the rectangular filtering matrix and

$$\mathbf{e}_{\text{rs}} = \mathbf{v}_{\text{rn}} \tag{18}$$

represents the residual noise. Therefore, the MSE criterion is:

$$\begin{aligned}J\left(\mathbf{H}\right) = \text{tr}\left[E\left(\mathbf{e}\mathbf{e}^H\right)\right] &= \text{tr}\left(\Phi_{\mathbf{x}_Q}\right) + \text{tr}\left(\mathbf{H}\Phi_{\mathbf{y}}\mathbf{H}^H\right) \\ &- \text{tr}\left(\mathbf{H}\Phi_{\mathbf{x}}\mathbf{I}_{\text{i}}^T\right) - \text{tr}\left(\mathbf{I}_{\text{i}}\Phi_{\mathbf{x}}\mathbf{H}^H\right),\end{aligned} \tag{19}$$

where

$$\mathbf{I}_{\text{i}} = \left[\begin{array}{cc}\mathbf{I}_Q & \mathbf{0}_{Q\times(M-Q)}\end{array}\right] \tag{20}$$

is the identity filtering matrix, with $\mathbf{I}_Q$ being the $Q \times Q$ identity matrix. Using the fact that $E\left(\mathbf{e}_{\text{ds}}\mathbf{e}_{\text{rs}}^H\right) = \mathbf{0}_{Q\times Q}$, $J\left(\mathbf{H}\right)$ can be expressed as the sum of two other MSEs, i.e.,

$$\begin{aligned}J\left(\mathbf{H}\right) &= \text{tr}\left[E\left(\mathbf{e}_{\text{ds}}\mathbf{e}_{\text{ds}}^H\right)\right] + \text{tr}\left[E\left(\mathbf{e}_{\text{rs}}\mathbf{e}_{\text{rs}}^H\right)\right] \\ &= J_{\text{ds}}\left(\mathbf{H}\right) + J_{\text{rs}}\left(\mathbf{H}\right),\end{aligned} \tag{21}$$

where

$$J_{\text{ds}}\left(\mathbf{H}\right) = \text{tr}\left(\Phi_{\mathbf{x}_Q}\right)\upsilon_{\text{sd}}\left(\mathbf{H}\right) \tag{22}$$

and

$$J_{\text{rs}}\left(\mathbf{H}\right) = \frac{\text{tr}\left(\Phi_{\mathbf{v}_Q}\right)}{\xi_{\text{nr}}\left(\mathbf{H}\right)}. \tag{23}$$

We deduce that

$$\begin{aligned}\frac{J_{\text{ds}}\left(\mathbf{H}\right)}{J_{\text{rs}}\left(\mathbf{H}\right)} &= \text{iSNR} \times \xi_{\text{nr}}\left(\mathbf{H}\right) \times \upsilon_{\text{sd}}\left(\mathbf{H}\right) \\ &= \text{oSNR}\left(\mathbf{H}\right) \times \xi_{\text{sr}}\left(\mathbf{H}\right) \times \upsilon_{\text{sd}}\left(\mathbf{H}\right).\end{aligned} \tag{24}$$

We observe how the MSEs are related to the different performance measures.

1.3.3 Optimal filters

Let $\lambda_{\max}$ be the maximum eigenvalue of the matrix $\Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{x}}$ with corresponding eigenvector $\mathbf{b}_{\max}$. It can be shown that the maximum SNR filtering matrix is given by [20]:

$$\mathbf{H}_{\max} = \begin{bmatrix} \beta_1 \mathbf{b}_{\max}^T \\ \beta_2 \mathbf{b}_{\max}^T \\ \vdots \\ \beta_Q \mathbf{b}_{\max}^T \end{bmatrix}, \tag{25}$$

where β_q, $q = 1, 2, \ldots, Q$ are arbitrary complex numbers with at least one of them different from 0. The corresponding output SNR is:

$$\text{oSNR}\left(\mathbf{H}_{\max}\right) = \lambda_{\max}. \tag{26}$$

The output SNR with the maximum SNR filtering matrix is always greater than or equal to the input SNR, i.e., $\text{oSNR}\left(\mathbf{H}_{\max}\right) \geq \text{iSNR}$. We also have $\text{oSNR}\left(\mathbf{H}\right) \leq \lambda_{\max}, \forall \mathbf{H}$. The best way to find the β_qs is by minimizing distortion. By substituting $\mathbf{H}_{\max}$ into the distortion-based MSE and minimizing with respect to the β_qs, we get

$$\begin{aligned} \mathbf{H}_{\max} &= \mathbf{I}_{\text{i}}\Phi_{\mathbf{x}}\frac{\mathbf{b}_{\max}\mathbf{b}_{\max}^H}{\lambda_{\max}} \\ &= \mathbf{I}_{\text{i}}\Phi_{\mathbf{v}}\mathbf{b}_{\max}\mathbf{b}_{\max}^H. \end{aligned} \tag{27}$$

If we differentiate the MSE criterion, $J\left(\mathbf{H}\right)$, with respect to $\mathbf{H}$ and equate the result to zero, we find the Wiener filtering matrix:

$$\begin{aligned} \mathbf{H}_{\text{W}} &= \mathbf{I}_{\text{i}}\Phi_{\mathbf{x}}\Phi_{\mathbf{y}}^{-1} \\ &= \mathbf{I}_{\text{i}}\left(\mathbf{I}_M - \Phi_{\mathbf{v}}\Phi_{\mathbf{y}}^{-1}\right), \end{aligned} \tag{28}$$

where $\mathbf{I}_M$ is the $M \times M$ identity matrix. The output SNR with the Wiener filtering matrix is always greater than or equal to the input SNR, i.e., $\text{oSNR}\left(\mathbf{H}_{\text{W}}\right) \geq \text{iSNR}$. Obviously, we have

$$\text{oSNR}\left(\mathbf{H}_{\text{W}}\right) \leq \text{oSNR}\left(\mathbf{H}_{\max}\right) \tag{29}$$

and, in general,

$$\upsilon_{\text{sd}}\left(\mathbf{H}_{\text{W}}\right) \leq \upsilon_{\text{sd}}\left(\mathbf{H}_{\max}\right). \tag{30}$$

To better compromise between noise reduction and speech distortion, we can minimize the speech distortion index with the constraint that the noise reduction factor is equal to a positive value that is greater than 1, i.e.,

$$\min_{\mathbf{H}} J_{\text{ds}}\left(\mathbf{H}\right) \quad \text{subject to} \quad J_{\text{rs}}\left(\mathbf{H}\right) = \beta \text{tr}\left(\Phi_{\mathbf{v}_Q}\right), \tag{31}$$

where $0 < \beta < 1$ to insure that we get some noise reduction. The previous optimization leads to the tradeoff filter:

$$\mathbf{H}_{\text{T},\mu} = \mathbf{I}_{\text{i}}\Phi_{\mathbf{x}}\left(\Phi_{\mathbf{x}} + \mu\Phi_{\mathbf{v}}\right)^{-1}, \tag{32}$$

where $\mu > 0$ is a Lagrange multiplier. The output SNR with the tradeoff filtering matrix is always greater than or equal to the input SNR, i.e., $\text{oSNR}\left(\mathbf{H}_{\text{T},\mu}\right) \geq \text{iSNR}$, $\forall \mu > 0$. Usually, μ is chosen in a heuristic way, so that for

- $\mu = 1$, $\mathbf{H}_{\text{T},1} = \mathbf{H}_{\text{W}}$, which is the Wiener filtering matrix;
- $\mu > 1$, results in a filtering matrix with low residual noise at the expense of high speech distortion (as compared to Wiener); and
- $\mu < 1$, results in a filtering matrix with high residual noise and low speech distortion (as compared to Wiener).

We should have for $\mu \geq 1$,

$$\text{oSNR}\left(\mathbf{H}_{\text{W}}\right) \leq \text{oSNR}\left(\mathbf{H}_{\text{T},\mu}\right) \leq \text{oSNR}\left(\mathbf{H}_{\max}\right), \tag{33}$$

$$\upsilon_{\text{sd}}\left(\mathbf{H}_{\text{W}}\right) \leq \upsilon_{\text{sd}}\left(\mathbf{H}_{\text{T},\mu}\right), \tag{34}$$

and for $\mu \leq 1$,

$$\text{oSNR}\left(\mathbf{H}_{\text{T},\mu}\right) \leq \text{oSNR}\left(\mathbf{H}_{\text{W}}\right) \leq \text{oSNR}\left(\mathbf{H}_{\max}\right), \tag{35}$$

$$\upsilon_{\text{sd}}\left(\mathbf{H}_{\text{T},\mu}\right) \leq \upsilon_{\text{sd}}\left(\mathbf{H}_{\text{W}}\right). \tag{36}$$

Another filter can be derived by just minimizing $J_{\text{ds}}\left(\mathbf{H}\right)$. We obtain the minimum distortion (MD) rectangular filtering matrix:

$$\mathbf{H}_{\text{MD}} = \mathbf{I}_{\text{i}}\Phi_{\mathbf{x}}\Phi_{\mathbf{x}}^{\dagger}, \tag{37}$$

where $\Phi_{\mathbf{x}}^{\dagger}$ is the pseudoinverse of $\Phi_{\mathbf{x}}$. If $\Phi_{\mathbf{x}}$ is a full-rank matrix, $\mathbf{H}_{\text{MD}}$ becomes the identity filter, $\mathbf{I}_{\text{i}}$, which does not affect the observations. The MD filter is very close to the well-known MVDR filter.

For $Q = P = 1$, it is possible to derive the MVDR filter. Indeed, by minimizing the variance of the filter's output, $\phi_z = \mathbf{h}^H\Phi_{\mathbf{y}}\mathbf{h}$, or the variance of the residual noise, $\phi_{v_{\text{rn}}} = \mathbf{h}^H\Phi_{\mathbf{v}}\mathbf{h}$, subject to the distortionless constraint, $\mathbf{h}^H\mathbf{d} = 1$, we easily get

$$\begin{aligned} \mathbf{h}_{\text{MVDR}} &= \frac{\Phi_{\mathbf{y}}^{-1}\mathbf{d}}{\mathbf{d}^H\Phi_{\mathbf{y}}^{-1}\mathbf{d}} \\ &= \frac{\Phi_{\mathbf{v}}^{-1}\mathbf{d}}{\mathbf{d}^H\Phi_{\mathbf{v}}^{-1}\mathbf{d}}. \end{aligned} \tag{38}$$

It can be checked that $J_{\text{ds}}\left(\mathbf{h}_{\text{MVDR}}\right) = 0$, proving that $\mathbf{h}_{\text{MVDR}}$ is distortionless.

It is also possible to derive the MVDR (square) filtering matrix for $Q = M$. Using the well-known eigenvalue decomposition [21], the speech correlation matrix can be diagonalized as:

$$\mathbf{Q}_{\mathbf{x}}^H\Phi_{\mathbf{x}}\mathbf{Q}_{\mathbf{x}} = \Lambda_{\mathbf{x}}, \tag{39}$$

where

$$\mathbf{Q}_{\mathbf{x}} = \left[\, \mathbf{q}_{\mathbf{x},1} \;\; \mathbf{q}_{\mathbf{x},2} \;\; \cdots \;\; \mathbf{q}_{\mathbf{x},M} \,\right] \tag{40}$$

is a unitary matrix, i.e., $\mathbf{Q}_{\mathbf{x}}^H\mathbf{Q}_{\mathbf{x}} = \mathbf{Q}_{\mathbf{x}}\mathbf{Q}_{\mathbf{x}}^H = \mathbf{I}_M$ and

$$\Lambda_{\mathbf{x}} = \text{diag}\left(\lambda_{\mathbf{x},1}, \lambda_{\mathbf{x},2}, \ldots, \lambda_{\mathbf{x},M}\right) \tag{41}$$

is a diagonal matrix. The orthonormal vectors $\mathbf{q}_{\mathbf{x},1}$, $\mathbf{q}_{\mathbf{x},2}, \ldots, \mathbf{q}_{\mathbf{x},M}$ are the eigenvectors corresponding, respectively, to the eigenvalues $\lambda_{\mathbf{x},1}, \lambda_{\mathbf{x},2}, \ldots, \lambda_{\mathbf{x},M}$ of the matrix $\Phi_{\mathbf{x}}$, where $\lambda_{\mathbf{x},1} \geq \lambda_{\mathbf{x},2} \geq \cdots \geq \lambda_{\mathbf{x},P} > 0$ and $\lambda_{\mathbf{x},P+1} = \lambda_{\mathbf{x},P+2} = \cdots = \lambda_{\mathbf{x},M} = 0$. Let

$$\mathbf{Q}_{\mathbf{x}} = \left[\mathbf{T}_{\mathbf{x}} \;\; \Xi_{\mathbf{x}} \right], \tag{42}$$

where the $M \times P$ matrix $\mathbf{T}_{\mathbf{x}}$ contains the eigenvectors corresponding to the nonzero eigenvalues of $\Phi_{\mathbf{x}}$ and the $M \times (M-P)$ matrix $\Xi_{\mathbf{x}}$ contains the eigenvectors corresponding to the null eigenvalues of $\Phi_{\mathbf{x}}$. It can be verified that

$$\mathbf{I}_M = \mathbf{T}_{\mathbf{x}}\mathbf{T}_{\mathbf{x}}^H + \Xi_{\mathbf{x}}\Xi_{\mathbf{x}}^H. \tag{43}$$

Notice that $\mathbf{T}_{\mathbf{x}}\mathbf{T}_{\mathbf{x}}^H$ and $\Xi_{\mathbf{x}}\Xi_{\mathbf{x}}^H$ are two orthogonal projection matrices of rank P and $M-P$, respectively. Hence, $\mathbf{T}_{\mathbf{x}}\mathbf{T}_{\mathbf{x}}^H$ is the orthogonal projector onto the speech subspace (where all the energy of the speech signal is concentrated), or the range of $\Phi_{\mathbf{x}}$ and $\Xi_{\mathbf{x}}\Xi_{\mathbf{x}}^H$ is the orthogonal projector onto the null subspace of $\Phi_{\mathbf{x}}$. Using Equation 43, we can write the speech vector as:

$$\begin{aligned} \mathbf{x} &= \mathbf{Q}_{\mathbf{x}}\mathbf{Q}_{\mathbf{x}}^H\mathbf{x} \\ &= \mathbf{T}_{\mathbf{x}}\mathbf{T}_{\mathbf{x}}^H\mathbf{x}. \end{aligned} \tag{44}$$

We deduce from Equation 44 that the distortionless constraint is:

$$\mathbf{H}\mathbf{T}_{\mathbf{x}} = \mathbf{T}_{\mathbf{x}}, \tag{45}$$

since, in this case, $\mathbf{H}\mathbf{x} = \mathbf{H}\mathbf{T}_{\mathbf{x}}\mathbf{T}_{\mathbf{x}}^H\mathbf{x} = \mathbf{T}_{\mathbf{x}}\mathbf{T}_{\mathbf{x}}^H\mathbf{x} = \mathbf{x}$. Now, from the criterion:

$$\min_{\mathbf{H}} \operatorname{tr}\left(\mathbf{H}\Phi_{\mathbf{v}}\mathbf{H}^H\right) \quad \text{subject to} \quad \mathbf{H}\mathbf{T}_{\mathbf{x}} = \mathbf{T}_{\mathbf{x}}, \tag{46}$$

we find the MVDR:

$$\mathbf{H}_{\text{MVDR}} = \mathbf{T}_{\mathbf{x}}\left(\mathbf{T}_{\mathbf{x}}^H\Phi_{\mathbf{v}}^{-1}\mathbf{T}_{\mathbf{x}}\right)^{-1}\mathbf{T}_{\mathbf{x}}^H\Phi_{\mathbf{v}}^{-1}. \tag{47}$$

Equation 47 can also be expressed as:

$$\mathbf{H}_{\text{MVDR}} = \mathbf{T}_{\mathbf{x}}\left(\mathbf{T}_{\mathbf{x}}^H\Phi_{\mathbf{y}}^{-1}\mathbf{T}_{\mathbf{x}}\right)^{-1}\mathbf{T}_{\mathbf{x}}^H\Phi_{\mathbf{y}}^{-1}. \tag{48}$$

Of course, for $P = M$, the MVDR filtering matrix simplifies to the identity matrix, i.e., $\mathbf{H}_{\text{MVDR}} = \mathbf{I}_M$. As a consequence, we can state that the higher the dimension of the null space of $\Phi_{\mathbf{x}}$ is, the more the MVDR is efficient in terms of noise reduction. The best scenario corresponds to $P = 1$. We can verify that $J_{\text{ds}}(\mathbf{H}_{\text{MVDR}}) = 0$.

1.4 Speech enhancement with the orthogonal decomposition of the speech signal vector

1.4.1 Principle

Another perspective for speech enhancement is to extract the desired signal vector, $\mathbf{x}_Q$, from $\mathbf{x}$. This way, the observation signal vector, $\mathbf{y}$, will be an explicit function of $\mathbf{x}_Q$. As a consequence, the objectives that we wish to achieve are much easier to handle.

In this section, we assume that the elements x_q, $q = 1, 2, \ldots, Q$ are not fully coherent, so that $\Phi_{\mathbf{x}_Q}$ is a full-rank matrix. To extract $\mathbf{x}_Q$ from $\mathbf{x}$, we need to decompose $\mathbf{x}$ into two orthogonal components: one correlated with (or is a linear transformation of) the desired signal vector and the other one orthogonal to $\mathbf{x}_Q$ and, hence, will be considered as an interference signal vector. Specifically, the vector $\mathbf{x}$ is decomposed into the following form [22,23]:

$$\begin{aligned} \mathbf{x} &= \Phi_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{x}_Q}^{-1}\mathbf{x}_Q + \mathbf{x}_{\text{i}} \\ &= \mathbf{x}_{\text{d}} + \mathbf{x}_{\text{i}}, \end{aligned} \tag{49}$$

where

$$\begin{aligned} \mathbf{x}_{\text{d}} &= \Phi_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{x}_Q}^{-1}\mathbf{x}_Q \\ &= \Gamma_{\mathbf{x}\mathbf{x}_Q}\mathbf{x}_Q \end{aligned} \tag{50}$$

is a linear transformation of the desired signal vector, $\Phi_{\mathbf{x}\mathbf{x}_Q} = E\left(\mathbf{x}\mathbf{x}_Q^H\right)$ is the cross-correlation matrix (of size $M \times Q$) between $\mathbf{x}$ and $\mathbf{x}_Q$, $\Gamma_{\mathbf{x}\mathbf{x}_Q} = \Phi_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{x}_Q}^{-1}$, and

$$\mathbf{x}_{\text{i}} = \mathbf{x} - \mathbf{x}_{\text{d}} \tag{51}$$

is the interference signal vector. It is easy to see that $\mathbf{x}_{\text{d}}$ and $\mathbf{x}_{\text{i}}$ are orthogonal[b], i.e.,

$$E\left(\mathbf{x}_{\text{d}}\mathbf{x}_{\text{i}}^H\right) = \mathbf{0}_{M\times M}. \tag{52}$$

We observe that the first Q elements of $\mathbf{x}_{\text{d}}$ and $\mathbf{x}_{\text{i}}$ are equal to $\mathbf{x}_Q$ and $\mathbf{0}_{Q\times 1}$, respectively. Now, we can express the observation signal vector as an explicit function of $\mathbf{x}_Q$, i.e.,

$$\mathbf{y} = \Gamma_{\mathbf{x}\mathbf{x}_Q}\mathbf{x}_Q + \mathbf{x}_{\text{i}} + \mathbf{v}. \tag{53}$$

With this approach, the estimator is:

$$\begin{aligned} \mathbf{z}' &= \mathbf{H}'\left[\mathbf{x}_{\text{d}} + \mathbf{x}_{\text{i}} + \mathbf{v}\right] \\ &= \mathbf{x}'_{\text{fd}} + \mathbf{x}'_{\text{ri}} + \mathbf{v}'_{\text{rn}}, \end{aligned} \tag{54}$$

where

$$\mathbf{H}' = \begin{bmatrix} \mathbf{h}_1'^H \\ \mathbf{h}_2'^H \\ \vdots \\ \mathbf{h}_Q'^H \end{bmatrix} \tag{55}$$

is a rectangular filtering matrix of size $Q \times M$, $\mathbf{h}'_q$, $q = 1, 2, \ldots, Q$ are complex-valued filters of length M, $\mathbf{x}'_{\text{fd}} = \mathbf{H}'\mathbf{x}_{\text{d}}$ is the filtered desired signal, $\mathbf{x}'_{\text{ri}} = \mathbf{H}'\mathbf{x}_{\text{i}}$ is the residual interference, and $\mathbf{v}'_{\text{rn}} = \mathbf{H}'\mathbf{v}$ is the residual noise. The correlation matrix of $\mathbf{z}'$ is then:

$$\Phi_{\mathbf{z}'} = \Phi_{\mathbf{x}'_{\text{fd}}} + \Phi_{\mathbf{x}'_{\text{ri}}} + \Phi_{\mathbf{v}'_{\text{rn}}}, \tag{56}$$

where $\Phi_{\mathbf{x}'_{\text{fd}}} = \mathbf{H}'\Phi_{\mathbf{x}_{\text{d}}}\mathbf{H}'^H$, with $\Phi_{\mathbf{x}_{\text{d}}} = \Gamma_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{x}_Q}\Gamma_{\mathbf{x}\mathbf{x}_Q}^H$ being the correlation matrix (whose rank is equal to Q) of $\mathbf{x}_{\text{d}}$, $\Phi_{\mathbf{x}'_{\text{ri}}} = \mathbf{H}'\Phi_{\mathbf{x}_{\text{i}}}\mathbf{H}'^H$, with $\Phi_{\mathbf{x}_{\text{i}}} = E\left(\mathbf{x}_{\text{i}}\mathbf{x}_{\text{i}}^H\right)$ being the correlation matrix of $\mathbf{x}_{\text{i}}$, and $\Phi_{\mathbf{v}'_{\text{rn}}} = \mathbf{H}'\Phi_{\mathbf{v}}\mathbf{H}'^H$.

1.4.2 Performance measures

The input SNR is identical to the definition given in Equation 9.

From Equation 56, we deduce the output SNR:

$$\begin{aligned}\text{oSNR}\left(\mathbf{H}'\right) &= \frac{\text{tr}\left(\Phi_{\mathbf{x}'_{\text{fd}}}\right)}{\text{tr}\left(\Phi_{\mathbf{x}'_{\text{ri}}}+\Phi_{\mathbf{v}'_{\text{rn}}}\right)}\\ &= \frac{\text{tr}\left(\mathbf{H}'\Gamma_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{x}_Q}\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\mathbf{H}'^H\right)}{\text{tr}\left(\mathbf{H}'\Phi_{\text{in}}\mathbf{H}'^H\right)},\end{aligned} \tag{57}$$

where

$$\Phi_{\text{in}} = \Phi_{\mathbf{x}_{\text{i}}} + \Phi_{\mathbf{v}} \tag{58}$$

is the correlation matrix of the interference-plus-noise. The obvious objective is to find an appropriate $\mathbf{H}'$ in such a way that $\text{oSNR}\left(\mathbf{H}'\right) \geq \text{iSNR}$.

The noise reduction factor is:

$$\xi_{\text{nr}}\left(\mathbf{H}'\right) = \frac{\text{tr}\left(\Phi_{\mathbf{v}_Q}\right)}{\text{tr}\left(\mathbf{H}'\Phi_{\text{in}}\mathbf{H}'^H\right)}. \tag{59}$$

A reasonable choice of $\mathbf{H}'$ should give a value of the noise reduction factor greater than 1, meaning that the noise and interference have been attenuated by the filter.

The speech reduction factor is defined as:

$$\xi_{\text{sr}}\left(\mathbf{H}'\right) = \frac{\text{tr}\left(\Phi_{\mathbf{x}_Q}\right)}{\text{tr}\left(\mathbf{H}'\Gamma_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{x}_Q}\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\mathbf{H}'^H\right)}. \tag{60}$$

A rectangular filtering matrix that does not affect the desired signal requires the constraint[c]:

$$\mathbf{H}'\Gamma_{\mathbf{x}\mathbf{x}_Q} = \mathbf{I}_Q. \tag{61}$$

Hence, $\xi_{\text{sr}}\left(\mathbf{H}'\right) = 1$ in the absence of (correlated) distortion and $\xi_{\text{sr}}\left(\mathbf{H}'\right) > 1$ in the presence of distortion.

Again, we have the fundamental relationship:

$$\frac{\text{oSNR}\left(\mathbf{H}'\right)}{\text{iSNR}} = \frac{\xi_{\text{nr}}\left(\mathbf{H}'\right)}{\xi_{\text{sr}}\left(\mathbf{H}'\right)}. \tag{62}$$

When no distortion occurs, the gain in SNR coincides with the noise reduction factor.

We can also quantify the distortion with the speech distortion index:

$$\upsilon_{\text{sd}}\left(\mathbf{H}'\right) = \frac{\text{tr}\left[\left(\mathbf{H}'\Gamma_{\mathbf{x}\mathbf{x}_Q}-\mathbf{I}_Q\right)\Phi_{\mathbf{x}_Q}\left(\mathbf{H}'\Gamma_{\mathbf{x}\mathbf{x}_Q}-\mathbf{I}_Q\right)^H\right]}{\text{tr}\left(\Phi_{\mathbf{x}_Q}\right)}. \tag{63}$$

The speech distortion index is always greater than or equal to 0 and should be upper bounded by 1 for optimal filtering matrices, which corresponds to the case where the filtering matrix is just a matrix of zeros; so the higher the value of $\upsilon_{\text{sd}}\left(\mathbf{H}'\right)$ is, the more the desired signal is distorted.

The error signal is:

$$\begin{aligned}\mathbf{e}' &= \mathbf{z}' - \mathbf{x}_Q\\ &= \mathbf{H}'\mathbf{y} - \mathbf{x}_Q.\end{aligned} \tag{64}$$

It can be written as the sum of two orthogonal error signal vectors:

$$\mathbf{e}' = \mathbf{e}'_{\text{ds}} + \mathbf{e}'_{\text{rs}}, \tag{65}$$

where

$$\begin{aligned}\mathbf{e}'_{\text{ds}} &= \mathbf{x}'_{\text{fd}} - \mathbf{x}_Q\\ &= \left(\mathbf{H}'\Gamma_{\mathbf{x}\mathbf{x}_Q} - \mathbf{I}_Q\right)\mathbf{x}_Q\end{aligned} \tag{66}$$

is the signal distortion due to the rectangular filtering matrix and

$$\begin{aligned}\mathbf{e}'_{\text{rs}} &= \mathbf{x}'_{\text{ri}} + \mathbf{v}'_{\text{rn}}\\ &= \mathbf{H}'\mathbf{x}_{\text{i}} + \mathbf{H}'\mathbf{v}\end{aligned} \tag{67}$$

represents the residual interference-plus-noise. Having defined the error signal, we can now write the MSE criterion:

$$\begin{aligned}J\left(\mathbf{H}'\right) &= \text{tr}\left[E\left(\mathbf{e}'\mathbf{e}'^H\right)\right]\\ &= \text{tr}\left(\Phi_{\mathbf{x}_Q}\right) + \text{tr}\left(\mathbf{H}'\Phi_{\mathbf{y}}\mathbf{H}'^H\right) - \text{tr}\left(\mathbf{H}'\Phi_{\mathbf{x}}\mathbf{I}_{\text{i}}^T\right)\\ &\quad - \text{tr}\left(\mathbf{I}_{\text{i}}\Phi_{\mathbf{x}}\mathbf{H}'^H\right)\\ &= J_{\text{ds}}\left(\mathbf{H}'\right) + J_{\text{rs}}\left(\mathbf{H}'\right),\end{aligned} \tag{68}$$

where

$$\begin{aligned}J_{\text{ds}}\left(\mathbf{H}'\right) &= \text{tr}\left(\Phi_{\mathbf{x}_Q}\right) + \text{tr}\left(\mathbf{H}'\Phi_{\mathbf{x}_{\text{d}}}\mathbf{H}'^H\right) - \text{tr}\left(\mathbf{H}'\Phi_{\mathbf{x}_{\text{d}}}\mathbf{I}_{\text{i}}^T\right)\\ &\quad - \text{tr}\left(\mathbf{I}_{\text{i}}\Phi_{\mathbf{x}_{\text{d}}}\mathbf{H}'^H\right)\end{aligned} \tag{69}$$

and

$$J_{\text{rs}}\left(\mathbf{H}'\right) = \mathbf{H}'\Phi_{\text{in}}\mathbf{H}'^H. \tag{70}$$

We deduce that

$$\begin{aligned}\frac{J_{\text{ds}}\left(\mathbf{H}'\right)}{J_{\text{rs}}\left(\mathbf{H}'\right)} &= \text{iSNR}\times\xi_{\text{nr}}\left(\mathbf{H}'\right)\times\upsilon_{\text{sd}}\left(\mathbf{H}'\right)\\ &= \text{oSNR}\left(\mathbf{H}'\right)\times\xi_{\text{sr}}\left(\mathbf{H}'\right)\times\upsilon_{\text{sd}}\left(\mathbf{H}'\right),\end{aligned} \tag{71}$$

showing how the MSEs are related to the different performance measures.

1.4.3 Optimal filters

Let $\lambda'_{\max}$ be the maximum eigenvalue of the matrix $\Phi_{\text{in}}^{-1}\Phi_{\mathbf{x}_{\text{d}}}$ with corresponding eigenvector $\mathbf{b}'_{\max}$. We easily find that the maximum SNR filtering matrix with minimum distortion is:

$$\begin{aligned}\mathbf{H}'_{\max} &= \mathbf{I}_{\text{i}}\Phi_{\mathbf{x}_{\text{d}}}\frac{\mathbf{b}'_{\max}\mathbf{b}'^H_{\max}}{\lambda'_{\max}}\\ &= \mathbf{I}_{\text{i}}\Phi_{\text{in}}\mathbf{b}'_{\max}\mathbf{b}'^H_{\max}\end{aligned} \tag{72}$$

and

$$\text{oSNR}\left(\mathbf{H}'_{\max}\right) = \lambda'_{\max}. \quad (73)$$

The output SNR with the maximum SNR filtering matrix is always greater than or equal to the input SNR, i.e., $\text{oSNR}\left(\mathbf{H}'_{\max}\right) \geq \text{iSNR}$. We also have $\text{oSNR}\left(\mathbf{H}'\right) \leq \lambda'_{\max}, \forall \mathbf{H}'$.

The minimization of the MSE criterion leads to the Wiener filtering matrix:

$$\begin{aligned}\mathbf{H}'_{\text{W}} &= \mathbf{I}_{\text{i}}\Phi_{\mathbf{x}}\Phi_{\mathbf{y}}^{-1} \\ &= \mathbf{H}_{\text{W}}, \end{aligned} \quad (74)$$

which is identical to the Wiener filter obtained with the classical approach. Even though the Wiener filter obtained with the two different approaches is the same, its evaluation with the performance measures is slightly different due the conceptual difference between the two methods. We always have $\text{oSNR}\left(\mathbf{H}'_{\text{W}}\right) \geq \text{iSNR}$.

We can rewrite the Wiener filtering matrix as:

$$\begin{aligned}\mathbf{H}'_{\text{W}} &= \left(\mathbf{I}_Q + \Phi_{\mathbf{x}_Q}\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\text{in}}^{-1}\Gamma_{\mathbf{x}\mathbf{x}_Q}\right)^{-1}\Phi_{\mathbf{x}_Q}\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\text{in}}^{-1} \\ &= \left(\Phi_{\mathbf{x}_Q}^{-1} + \Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\text{in}}^{-1}\Gamma_{\mathbf{x}\mathbf{x}_Q}\right)^{-1}\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\text{in}}^{-1}. \end{aligned} \quad (75)$$

This form is interesting because it shows an obvious link with some other optimal filtering matrices as it will be verified later.

Another way to express the Wiener filter is:

$$\begin{aligned}\mathbf{H}'_{\text{W}} &= \mathbf{I}_{\text{i}}\Gamma_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{x}_Q}\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{y}}^{-1} \\ &= \mathbf{I}_{\text{i}}\left(\mathbf{I}_M - \Phi_{\text{in}}\Phi_{\mathbf{y}}^{-1}\right). \end{aligned} \quad (76)$$

The MVDR[d] rectangular filtering matrix is obtained by minimizing the MSE of the residual interference-plus-noise, $J_{\text{rs}}\left(\mathbf{H}'\right)$, subject to the constraint that the desired signal vector is not distorted. Mathematically, this is equivalent to:

$$\min_{\mathbf{H}'} \text{tr}\left(\mathbf{H}'\Phi_{\text{in}}\mathbf{H}'^H\right) \quad \text{subject to} \quad \mathbf{H}\Gamma_{\mathbf{x}\mathbf{x}_Q} = \mathbf{I}_Q. \quad (77)$$

The solution to the above optimization problem is:

$$\mathbf{H}'_{\text{MVDR}} = \left(\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\text{in}}^{-1}\Gamma_{\mathbf{x}\mathbf{x}_Q}\right)^{-1}\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\text{in}}^{-1}, \quad (78)$$

which is interesting to compare to $\mathbf{H}'_{\text{W}}$ [Equation 75]. We can rewrite the MVDR as:

$$\mathbf{H}'_{\text{MVDR}} = \left(\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{y}}^{-1}\Gamma_{\mathbf{x}\mathbf{x}_Q}\right)^{-1}\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{y}}^{-1}. \quad (79)$$

We should always have

$$\text{oSNR}\left(\mathbf{H}'_{\text{MVDR}}\right) \leq \text{oSNR}\left(\mathbf{H}'_{\text{W}}\right) \leq \text{oSNR}\left(\mathbf{H}'_{\max}\right). \quad (80)$$

By minimizing the speech distortion index with the constraint that the noise reduction factor is equal to a positive value that is greater than 1, we find the tradeoff filtering matrix:

$$\mathbf{H}'_{\text{T},\mu'} = \Phi_{\mathbf{x}_Q}\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\left(\Gamma_{\mathbf{x}\mathbf{x}_Q}\Phi_{\mathbf{x}_Q}\Gamma^H_{\mathbf{x}\mathbf{x}_Q} + \mu'\Phi_{\text{in}}\right)^{-1}, \quad (81)$$

which can be rewritten as:

$$\mathbf{H}'_{\text{T},\mu'} = \left(\mu'\Phi_{\mathbf{x}_Q}^{-1} + \Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\text{in}}^{-1}\Gamma_{\mathbf{x}\mathbf{x}_Q}\right)^{-1}\Gamma^H_{\mathbf{x}\mathbf{x}_Q}\Phi_{\text{in}}^{-1}, \quad (82)$$

where $\mu' \geq 0$ is a Lagrange multiplier. Usually, μ' is chosen in an *ad hoc* way, so that for

- $\mu' = 1$, $\mathbf{H}'_{\text{T},1} = \mathbf{H}'_{\text{W}}$, which is the Wiener filtering matrix;
- $\mu' = 0$ [from Equation 82], $\mathbf{H}'_{\text{T},0} = \mathbf{H}'_{\text{MVDR}}$, which is the MVDR filtering matrix;
- $\mu' > 1$, results in a filtering matrix with low residual noise (as compared to Wiener) at the expense of high speech distortion; and
- $\mu' < 1$, results in a filtering matrix with high residual noise and low speech distortion (as compared to Wiener).

We always have $\text{oSNR}\left(\mathbf{H}'_{\text{T},\mu'}\right) \geq \text{iSNR}$, $\forall \mu' \geq 0$. We should also have for $\mu' \geq 1$,

$$\begin{aligned}\text{oSNR}\left(\mathbf{H}'_{\text{MVDR}}\right) &\leq \text{oSNR}\left(\mathbf{H}'_{\text{W}}\right) \leq \text{oSNR}\left(\mathbf{H}'_{\text{T},\mu'}\right) \\ &\leq \text{oSNR}\left(\mathbf{H}'_{\max}\right) \end{aligned} \quad (83)$$

and for $\mu' \leq 1$,

$$\begin{aligned}\text{oSNR}\left(\mathbf{H}'_{\text{MVDR}}\right) &\leq \text{oSNR}\left(\mathbf{H}'_{\text{T},\mu'}\right) \leq \text{oSNR}\left(\mathbf{H}'_{\text{W}}\right) \\ &\leq \text{oSNR}\left(\mathbf{H}'_{\max}\right). \end{aligned} \quad (84)$$

The case $Q = M$ is interesting because for both approaches, performance measures and optimal square filtering matrices are identical. We can draw the same conclusions for the case $Q = P = 1$.

1.5 Application examples

In this section, we show how the two approaches can be applied to different applications of speech enhancement.

1.5.1 Single-channel noise reduction in the time domain

The single-channel noise reduction problem in the time domain consists of recovering the desired signal (or clean speech) $x(t)$, t being the discrete-time index, of zero mean from the noisy observation (microphone signal) [1]:

$$y(t) = x(t) + v(t), \quad (85)$$

where $v(t)$, assumed to be a zero-mean random process, is the unwanted additive noise that can be either white or colored but is uncorrelated with $x(t)$.

The signal model given in Equation 85 can be put into a vector form by considering the L most recent successive time samples, i.e.,

$$\begin{aligned}\mathbf{y}(t) &= \left[\, y(t)\ y(t-1)\ \cdots\ y(t-L+1)\,\right]^T \\ &= \mathbf{x}(t)+\mathbf{v}(t),\end{aligned} \tag{86}$$

where $\mathbf{x}(t)$ and $\mathbf{v}(t)$ are defined in a similar way to $\mathbf{y}(t)$. We define the desired signal vector as:

$$\mathbf{x}_Q(t)=\left[\, x(t)\ x(t-1)\ \cdots\ x(t-Q+1)\,\right]^T, \tag{87}$$

that we can estimate from $\mathbf{y}(t)$ with either of the two methods. Estimating the desired signal using conventional, rectangular filters was considered in [24], while [22] considers the orthogonal decomposition approach. Simulation results showing the performance of the two filtering methods for single-channel noise reduction are also found in [22,24].

1.5.2 Single-channel noise reduction in the time-frequency domain

Using the short-time Fourier transform (STFT), Equation 85 can be rewritten in the time-frequency domain as [13,25]:

$$Y(k,n)=X(k,n)+V(k,n), \tag{88}$$

where the zero-mean complex random variables $Y(k,n)$, $X(k,n)$, and $V(k,n)$ are the STFTs of $y(t)$, $x(t)$, and $v(t)$, respectively, at frequency bin $k\in\{0,1,\ldots,K-1\}$ and time frame n. Here, the sample $X(k,n)$ is the desired signal.

The simplest way to estimate $X(k,n)$ is by applying a positive gain to $Y(k,n)$ with the conventional approach. However, the noise reduction performance may be limited.

A more general approach to estimate the desired signal is by filtering the observation signal vector of length P [25]:

$$\mathbf{y}(k,n)=[Y(k,n)\ \ Y(k,n-1)\ \ \cdots\ \ Y(k,n-P+1)]^T, \tag{89}$$

and using the orthogonal decomposition to extract $X(k,n)$ from $\mathbf{x}(k,n)$, which is defined in a similar way to $\mathbf{y}(k,n)$. Thanks to this approach, the non-negligible interframe correlation is taken into account, which is not the case when just a gain is used. As a consequence, we can better compromise between noise reduction and speech distortion.

The STFT-based filtering methods for single-channel noise reduction was considered in, e.g., [26-28], where experimental results can also be found.

1.5.3 Multichannel noise reduction in the time domain

In the multichannel scenario, we have a microphone array with M sensors that captures a convolved source signal in some noise field. In the time domain, the received signals are expressed as [29,30]:

$$\begin{aligned}y_m(t) &= g_m(t)*s(t)+v_m(t) \\ &= x_m(t)+v_m(t),\ m=1,2,\ldots,M,\end{aligned} \tag{90}$$

where $g_m(t)$ is the acoustic impulse response from the unknown speech source, $s(t)$, location to the mth microphone, $*$ stands for linear convolution, and $x_m(t)$ and $v_m(t)$ are, respectively, the convolved speech and additive noise at microphone m. We assume that the signals $x_m(t)=g_m(t)*s(t)$ and $v_m(t)$ are uncorrelated, zero mean, real, and broadband. By definition, $x_m(t)$ is coherent across the array. The noise signals, $v_m(t)$, are typically only partially coherent across the array.

By processing the data by blocks of L time samples, the signal model given in Equation 90 can be put into a vector form as:

$$\mathbf{y}_m(t)=\mathbf{x}_m(t)+\mathbf{v}_m(t),\ m=1,2,\ldots,M, \tag{91}$$

where

$$\mathbf{y}_m(t)=\left[\, y_m(t)\ y_m(t-1)\ \cdots\ y_m(t-L+1)\,\right]^T \tag{92}$$

is a vector of length L, and $\mathbf{x}_m(t)$ and $\mathbf{v}_m(t)$ are defined similarly to $\mathbf{y}_m(t)$. It is more convenient to concatenate the M vectors $\mathbf{y}_m(t),\ m=1,2,\ldots,M$ together as:

$$\begin{aligned}\underline{\mathbf{y}}(t) &= \left[\, \mathbf{y}_1^T(t)\ \mathbf{y}_2^T(t)\ \cdots\ \mathbf{y}_M^T(t)\,\right]^T \\ &= \underline{\mathbf{x}}(t)+\underline{\mathbf{v}}(t),\end{aligned} \tag{93}$$

where vectors $\underline{\mathbf{x}}(t)$ and $\underline{\mathbf{v}}(t)$ of length ML are defined in a similar way to $\underline{\mathbf{y}}(t)$.

We consider $\mathbf{x}_1(t)$ as the desired signal vector. Our problem then may be stated as follows: given $\underline{\mathbf{y}}(t)$, our aim is to preserve $\mathbf{x}_1(t)$ while minimizing the contribution of $\underline{\mathbf{v}}(t)$. Both approaches can be used but the one based on the orthogonal decomposition is more appropriate since it will better exploit the correlation among the convolved speech signals at the microphones for noise reduction. The orthogonal decomposition approach for multichannel noise reduction was considered in, e.g., [31,32], where experimental results can also be found.

1.5.4 Multichannel noise reduction in the frequency domain

In the frequency domain, at the frequency index f, Equation 90 can be expressed as:

$$\begin{aligned}Y_m(f) &= G_m(f)S(f)+V_m(f) \\ &= X_m(f)+V_m(f),\ m=1,2,\ldots,M,\end{aligned} \tag{94}$$

where $Y_m(f)$, $G_m(f)$, $S(f)$, $V_m(f)$, and $X_m(f)$ are the frequency-domain representations of $y_m(t)$, $g_m(t)$, $s(t)$, $v_m(t)$, and $x_m(t)$, respectively.

It is more convenient to write the M frequency-domain microphone signals in a vector notation:

$$\begin{aligned}\mathbf{y}(f) &= \mathbf{g}(f)S(f) + \mathbf{v}(f)\\ &= \mathbf{x}(f) + \mathbf{v}(f)\\ &= \mathbf{d}(f)X_1(f) + \mathbf{v}(f),\end{aligned} \tag{95}$$

where

$$\begin{aligned}\mathbf{y}(f) &= \left[\, Y_1(f)\ Y_2(f)\ \cdots\ Y_M(f)\,\right]^T,\\ \mathbf{x}(f) &= \left[\, X_1(f)\ X_2(f)\ \cdots\ X_M(f)\,\right]^T\\ &= S(f)\mathbf{g}(f),\\ \mathbf{g}(f) &= \left[\, G_1(f)\ G_2(f)\ \cdots\ G_M(f)\,\right]^T,\\ \mathbf{v}(f) &= \left[\, V_1(f)\ V_2(f)\ \cdots\ V_M(f)\,\right]^T,\end{aligned}$$

and

$$\begin{aligned}\mathbf{d}(f) &= \left[\, 1\ \frac{G_2(f)}{G_1(f)}\ \cdots\ \frac{G_M(f)}{G_1(f)}\,\right]^T\\ &= \frac{\mathbf{g}(f)}{G_1(f)}.\end{aligned} \tag{96}$$

Expression in Equation 95 depends explicitly on the desired signal, $X_1(f)$, that we want to estimate from $\mathbf{y}(f)$.

There is another interesting way to write Equation 95. First, it is easy to see that

$$X_m(f) = \gamma_{X_mX_1}(f)X_1(f),\ m = 1, 2, \ldots, M, \tag{97}$$

where

$$\begin{aligned}\gamma_{X_mX_1}(f) &= \frac{E\left[X_m(f)X_1^*(f)\right]}{E\left[\left|X_1(f)\right|^2\right]}\\ &= \frac{G_m(f)}{G_1(f)},\ m = 1, 2, \ldots, M\end{aligned} \tag{98}$$

is the partially normalized [with respect to $X_1(f)$] coherence function between $X_m(f)$ and $X_1(f)$. Using Equation 97, we can rewrite Equation 95 as:

$$\mathbf{y}(f) = \boldsymbol{\gamma}_{\mathbf{x}X_1}(f)X_1(f) + \mathbf{v}(f), \tag{99}$$

where

$$\begin{aligned}\boldsymbol{\gamma}_{\mathbf{x}X_1}(f) &= \left[\, 1\ \gamma_{X_2X_1}(f)\ \cdots\ \gamma_{X_MX_1}(f)\,\right]^T\\ &= \frac{E\left[\mathbf{x}(f)X_1^*(f)\right]}{E\left[\left|X_1(f)\right|^2\right]}\\ &= \mathbf{d}(f)\end{aligned} \tag{100}$$

is the partially normalized [with respect to $X_1(f)$] coherence vector (of length M) between $\mathbf{x}(f)$ and $X_1(f)$. This shows that the two approaches for noise reduction are identical. More details on multichannel noise reduction in the frequency domain as well as experimental results can be found in [25,33].

1.5.5 Binaural noise reduction

Binaural noise reduction [34] consists of the estimation of the received speech signal at two different microphones with a sensor array of M microphones. One estimate is for the left ear and the other for the right ear. This way and thanks to our binaural hearing system, we will be able to localize the speech source in the space. In the frequency domain, we can estimate, for example, $X_1(f)$ and $X_2(f)$. As explained above, the two methods are the same. In the time domain, we can estimate, for example, $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$. The method based on the orthogonal decomposition is more appropriate since it may distort less the signals. Distortion in binaural noise reduction is problematic since it may affect the cues for localization and separation. Experimental results and further theoretical details on binaural noise reduction using the approaches mentioned herein are found in [35,36].

2 Conclusions

In this paper, we have given a brief overview of linear filtering methods for speech enhancement based on two approaches: a so-called conventional approach and an approach based on the orthogonal decomposition. In the context of these two different approaches, various optimal filters (e.g., MVDR, maximum SNR, and Wiener filters) have been derived and their properties in terms of different performance measures have been assessed and compared. These performance measures, simply put, quantify the properties of the filters and approaches in terms of noise reduction and speech distortion and show how they offer different tradeoffs between the two. We have also demonstrated how the approaches can be applied in various speech enhancement contexts, including single- and multichannel enhancement in both the time and frequency domains and in binaural noise reduction.

Endnotes

[a]The upper bound comes from the fact that this distortion is obtained when the filtering matrix only contains zeros, which should be the maximum expected distortion.

[b]It is legitimate to consider $\mathbf{x}_\mathrm{i}$ as an interference, since the desired signal is entirely in $\mathbf{x}_\mathrm{d}$, and $\mathbf{x}_\mathrm{i}$ and $\mathbf{x}_\mathrm{d}$ are uncorrelated.

[c]Here, the distortionless constraint is in the sense that we can perfectly recover the desired signal vector, even though the residual interference can add some uncorrelated distortion to the desired signal.

[d]We use the terminology MVDR because we can completely extract the desired signal with this filter.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors contributed equally to this work. All authors read and approved the final manuscript.

Acknowledgements
This research was funded by the Villum Foundation and the Danish Council for Independent Research, Grant ID: DFF – 1337-00084.

Author details
[1]Audio Analysis Lab, AD:MT, Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark. [2]INRS-EMT, University of Quebec, 800 De La Gauchetiere Ouest, H5A 1K6 Montreal, Quebec, Canada. [3]Northwestern Polytechnical University, 127 Youyi West Rd, 710072 Xi'an, Shaanxi, China.

Received: 2 June 2014 Accepted: 31 October 2014
Published: 13 November 2014

References

1. J Chen, J Benesty, Y Huang, S Doclo, New insights into the noise reduction Wiener filter. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1218–1234 (2006). doi:10.1109/TSA.2005.860851
2. S Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **27**(2), 113–120 (1979). doi:10.1109/TASSP.1979.1163209
3. Y Ephraim, HL Van Trees, A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. **3**(4), 251–266 (1995). doi:10.1109/89.397090
4. SH Jensen, PC Hansen, SD Hansen, JA Sørensen, Reduction of broad-band noise in speech by truncated QSVD. IEEE Trans. Speech Audio Process. **3**(6), 439–448 (1995). doi:10.1109/89.482211
5. RJ McAulay, ML Malpass, Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust. Speech Signal Process. **28**(2), 137–145 (1980). doi:10.1109/TASSP.1980.1163394
6. Y Ephraim, D Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **33**(2), 443–445 (1985). doi:10.1109/TASSP.1985.1164550
7. S Srinivasan, J Samuelsson, WB Kleijn, Codebook-based bayesian speech enhancement for nonstationary environments. IEEE Trans. Audio Speech Lang. Process. **15**(2), 441–452 (2007). doi:10.1109/TASL.2006.881696
8. LR Rabiner, MR Sambur, An algorithm for determining the endpoints of isolated utterances. Bell Syst. Tech. J. **54**(2), 297–315 (1975)
9. R Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001). doi:10.1109/89.928915
10. J Freudenberger, S Stenzel, B Venditti, in *Proc. IEEE Workshop Statist. Signal Process*. A noise PSD and cross-PSD estimation for two-microphone speech enhancement systems, (2009), pp. 709–712. doi:10.1109/SSP.2009.5278478
11. T Lotter, P Vary, Dual-channel speech enhancement by superdirective beamforming. EURASIP J. Appl. Signal Process. **2006**(1), 1–14 (2006). doi:10.1155/ASP/2006/63297
12. RC Hendriks, T Gerkmann, Noise correlation matrix estimation for multi-microphone speech enhancement. IEEE Trans. Audio Speech Lang. Process. **20**(1), 223–233 (2012). doi:10.1109/TASL.2011.2159711
13. J Benesty, J Chen, Y Huang, I Cohen, *Noise Reduction in Speech Processing*. (Springer, Berlin, 2009)
14. P Loizou, *Speech Enhancement: Theory and Practice*. (CRC Press, Boca Raton, 2007)
15. P Vary, R Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. (John Wiley & Sons Ltd, Chichester, England, 2006)
16. S Doclo, M Moonen, GSVD-based optimal filtering for single and multimicrophone speech enhancement. IEEE Trans. Signal Process. **50**(9), 2230–2244 (2002). doi:10.1109/TSP.2002.801937
17. M Dendrinos, S Bakamidis, G Carayannis, Speech enhancement from noise: a regenerative approach. Speech Commun. **10**(1), 45–57 (1991). doi:10.1016/0167-6393(91)90027-Q
18. Y Hu, PC Loizou, A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans. Speech Audio Process. **11**, 334–341 (2003). doi:10.1109/TSA.2003.814458
19. K Hermus, P Wambacq, HV Hamme, A review of signal subspace speech enhancement and its application to noise robust speech recognition. EURASIP J. Adv. Signal Process. **2007**(1–15) (2007). Article ID 45821
20. J Benesty, J Chen, *Optimal Time-Domain Noise Reduction Filters - A Theoretical Study*, 1st edn. Springer Briefs in Electrical and Computer Engineering. (Springer, Berlin, 2011)
21. GH Golub, CF van Loan, *Matrix Computations*, 3rd edn. (The John Hopkins University Press, Baltimore, 1996)
22. J Benesty, J Chen, Y Huang, T Gaensler, Time-domain noise reduction based on an orthogonal decomposition for desired signal extraction. J. Acoust. Soc. Am. **132**(1), 452–464 (2012). doi:10.1121/1.4726071
23. T Long, J Chen, J Benesty, Z Zhang, Single-channel noise reduction using optimal rectangular filtering matrices. J. Acoust. Soc. Am. **133**, 1090–1101 (2013). doi:10.1121/1.4773269
24. JR Jensen, J Benesty, MG Christensen, J Chen, A class of optimal rectangular filtering matrices for single-channel signal enhancement in the time domain. IEEE Trans. Audio Speech Lang. Process. **21**(12), 2595–2606 (2013). doi:10.1109/TASL.2013.2280215
25. J Benesty, J Chen, E Habets, *Speech Enhancement In the STFT Domain*, Springer Briefs in Electrical and Computer Engineering. (Springer, Berlin, 2011)
26. J Benesty, Y Huang, A perspective on single-channel frequency-domain speech enhancement. Synth. Lect. Speech Audio Process. **8**(1), 1–101 (2011)
27. H Huang, L Zhao, J Chen, J Benesty, A minimum variance distortionless response filter based on the bifrequency spectrum for single-channel noise reduction. Digit. Signal Process. **33**(0), 169–179 (2014). doi:10.1016/j.dsp.2014.06.008
28. J Chen, J Benesty, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*. Single-channel noise reduction in the STFT domain based on the bifrequency spectrum, (2012), pp. 97–100. doi:10.1109/ICASSP.2012.6287826
29. J Benesty, Y Huang, J Chen, *Microphone Array Signal Processing*, vol. 1. (Springer, Berlin, 2008)
30. M Brandstein, D Ward (eds.), *Microphone Arrays - Signal Processing Techniques and Applications*. (Springer, Berlin, Germany, 2001)
31. J Benesty, M Souden, J Chen, A perspective on multichannel noise reduction in the time domain. Appl. Acoust. **74**(3), 343–355 (2013). doi:10.1016/j.apacoust.2012.08.002
32. JR Jensen, MG Christensen, J Benesty, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*. Multichannel signal enhancement using non-causal, time-domain filters, (2013), pp. 7274–7278. doi:10.1109/ICASSP.2013.6639075
33. M Souden, J Benesty, S Affes, On optimal frequency-domain multichannel linear filtering for noise reduction. IEEE Trans. Audio Speech Lang. Process. **18**(2), 260–276 (2010). doi:10.1109/TASL.2009.2025790
34. B Kollmeier, J Peissig, V Hohmann, Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain. Scand. Audiol. Suppl. **38**, 28–38 (1993)
35. J Benesty, J Chen, Y Huang, Binaural noise reduction in the time domain with a stereo setup. IEEE Trans. Audio Speech Lang. Process. **19**(8), 2260–2272 (2011). doi:10.1109/TASL.2011.2119313
36. J Chen, J Benesty, in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust*. A time-domain widely linear MVDR filter for binaural noise reduction, (2011), pp. 105–108. doi:10.1109/ASPAA.2011.6082262

doi:10.1186/1687-6180-2014-162
Cite this article as: Benesty *et al.*: **A brief overview of speech enhancement with linear filtering.** *EURASIP Journal on Advances in Signal Processing* 2014 **2014**:162.