Denoising Through Source Separation and Minimum Tracking

Sriram Srinivasan, Mattias Nilsson and W. Bastiaan Kleijn

Department of Signals, Sensors and Systems,
KTH (Royal Institute of Technology), Stockholm, Sweden
{sriram.srinivasan,mattias.nilsson,bastiaan.kleijn}@s3.kth.se

Abstract
In this paper, we develop a multi-channel noise reduction algorithm based on blind source separation (BSS). In contrast to general BSS algorithms that attempt to recover all the signals, we explicitly estimate only the speech signal. By tracking the minimum of the spectral density of the microphone signals, noise-only segments are identified. The coefficients of the unmixing matrix that are necessary to separate the speech are identified from these segments through the optimization of an appropriate energy criterion. Since the proposed method explicitly estimates the speech signal from the noisy mixture, it does not suffer from the permutation problem that is typical to conventional BSS techniques. The method is applicable to both instantaneous and convolutive mixtures and achieves the separation in a single step, without the need for iterations. Experimental results show superior performance compared to a general BSS algorithm.

1. Introduction
Acoustic background noise reduction is a relevant and challenging problem. Apart from reducing listener fatigue, noise reduction is an important front-end for speech coders, speech recognition systems, speaker identification systems etc. The bulk of the research in this area has been on single-channel enhancement, where a single microphone provides the noisy signal as input. Methods in this category include Wiener filtering and spectral subtraction [1], those using different statistical models on the speech and noise signals [2], and methods that use prior knowledge of the speech and noise processes [3].

Single-channel methods have been popular due to size and cost factors. With decreasing hardware costs and increasing computational power, interest in multi-channel speech enhancement methods has grown over the last years, for e.g., methods based on adaptive beamforming [4]. A related problem in the multi-microphone context is blind source separation (BSS) which is the separation of two or more signals that have been mixed through a channel. A typical example is the cocktail party problem where several speakers are active simultaneously.

Much progress has been made in this field and several algorithms have been developed [5,6]. These methods are blind in the sense that they do not require information about the sensor array geometry or direction of arrival of the signals unlike beamformers, which require this information.

The method presented in [6] is a general BSS algorithm applicable for both instantaneous and convolutive speech mixtures. For convolutive mixtures, the method operates in the frequency domain, and relies on the nonstationarity or nonwhiteness of the sources. The forward (the mixing matrix) or the backward (the unmixing matrix) model is estimated through simultaneous diagonalization of the cross power spectrum of the microphone signals at multiple time instants. Separation of noisy mixtures is also addressed.

In this paper, we consider background noise reduction as a convolutive BSS problem as shown in Fig. 1. We consider the case of a single speaker in the presence of one or more point noise sources and for simplicity show only one noise source in the figure.

![Schematic diagram of the linear convolutive mixing process.](image)

Figure 1: Schematic diagram of the linear convolutive mixing process. \( a_{ij} \) denotes the filter from the \( i \)th source to the \( j \)th sensor. The noisy signals \( y_1 \) and \( y_2 \) are observed at the microphones.

While the BSS approach of [6] can be applied to solve this problem, better performance can be achieved by exploiting the knowledge that only one of the signals is of interest in this case and that the noise signal can be discarded. The algorithm is developed in the frequency domain, which allows us to easily handle the convolutive mixing as instantaneous mixing in each bin. For a given frequency bin, from segments where the speech signal has little or no energy, the mixing parameters relevant to recover the speech signal are identified. These segments are determined by tracking the minimum of the spectral densities of the microphone signals. By tailoring the source separation to this specific case, we show that better performance can be achieved compared to [6] at a significantly lower computational complexity. While the algorithm of [6] requires several iterations for convergence, the proposed method converges in a single step.

2. Signal model
We consider the problem of acoustic background noise reduction using two microphones. The denoising problem can be cast in a BSS framework:

\[
y(t) = A x(t),
\]

where \( y(t) = [y_1(t) \ y_2(t)]^T \) are the observed microphone signals sampled at time instant \( t \), \( x(t) = [s(t) \ n(t)]^T \) with \( s(t) \) denoting the clean speech signal and \( n(t) \) denoting the...
noise source and \( A \) is the \( 2 \times 2 \) mixing matrix given by

\[
A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.
\]

The above model is referred to as the instantaneous mixing model. Without loss of generality, we assume that the signals \( s(t) \) and \( n(t) \) are each normalized to have unit variance. The coefficients of the mixing matrix \( A \) determine the signal-to-noise ratio (SNR) of the observed microphone signals. The instantaneous mixing problem can be solved only up to a permutation and scaling of the signals.

In practice, for acoustic signals, the mixing is better described by a convolutive model as given by the following \( Q + 1 \) tap mixing system:

\[
y(t) = \sum_{q=0}^{Q} A_q x(t-q),
\]

where \( A_q \) is a \( 2 \times 2 \) matrix for each \( q \). For large frame lengths, the convolution can be assumed to be circular so that it can be expressed as a product in the frequency domain. We thus have

\[
Y(\omega) = A(\omega)X(\omega),
\]

where

\[
A(\omega) = \begin{bmatrix} Y_1(\omega) \\ Y_2(\omega) \end{bmatrix}, \text{ } S(\omega) = \begin{bmatrix} S(\omega) \\ N(\omega) \end{bmatrix}
\]

and

\[
Y_i(\omega) = \sum_{t=0}^{T-1} y_i(t)e^{-j\omega t}, \text{ } i = 1, 2.
\]

\( S(\omega) \) and \( N(\omega) \) are obtained similarly from \( s(t) \) and \( n(t) \) respectively. \( A(\omega) \) corresponds to the frequency response of the mixing filters. The convolutive problem can thus be treated as an instantaneous mixing problem in each frequency bin. We observe that this leads to a permutation problem in each frequency bin, which unless resolved consistently for all frequency bins results in poor performance. Similar to the instantaneous case, the mixing matrix \( A(\omega) \) can be written as

\[
A(\omega) = \begin{bmatrix} a_{11}(\omega) & a_{12}(\omega) \\ a_{21}(\omega) & a_{22}(\omega) \end{bmatrix}
\]

(5), we wish to determine the mixing coefficients necessary to obtain the speech signal. In section 3.1, we describe the proposed approach that uses an estimate of the noise spectral density. In section 3.2, we describe a procedure to estimate the noise spectral density.

### 3.1. Denoising using the noise spectral density

In BSS, it is common to estimate the unmixing matrix so that an estimate of the original signals can be obtained as

\[
Z(\omega) = (Z_1(\omega) \ Z_2(\omega))^T = W(\omega) Y(\omega),
\]

where we assume \( W(\omega) \) to have the form

\[
W(\omega) = \begin{bmatrix} \alpha(\omega) & 1 - \alpha(\omega) \\ 1 - \beta(\omega) & \beta(\omega) \end{bmatrix}
\]

so that

\[
Z_1(\omega) = \alpha(\omega) Y_1(\omega) + (1 - \alpha(\omega)) Y_2(\omega).
\]

We wish to determine \( \alpha(\omega) \) such that \( Z_1(\omega) \) contains an estimate of the speech signal. As mentioned earlier, in contrast to other source separation algorithms, we do not wish to obtain \( \beta(\omega) \) since we are not interested in recovering the noise signal.

Let \( \hat{Y}(\omega) \) and \( \hat{Z}(\omega) \) denote the values of \( Y(\omega) \) and \( Z(\omega) \) during regions of speech absence. Thus \( Z_1(\omega) \) consists of noise alone. We can thus formulate an error criterion to be minimized as

\[
\eta(\omega) = E[\hat{Z}_1(\omega) \hat{Z}_1^*(\omega)]
\]

where the superscript * denotes complex conjugate transpose and \( E \) is the statistical expectation operator. Minimizing \( \eta(\omega) \) corresponds to minimizing the noise energy in noise-only segments. An estimate \( \hat{\alpha}(\omega) \) of \( \alpha(\omega) \) can be obtained as

\[
\hat{\alpha}(\omega) = \arg \min_{\alpha(\omega)} E[\hat{Z}_1(\omega) \hat{Z}_1^*(\omega)].
\]

While the above energy minimization appears similar to adaptive noise cancellation (ANC), in contrast to ANC which optimizes for \( w_1(\omega) \) in \( Z_1(\omega) = Y_1(\omega) + w_1(\omega) Y_2(\omega) \), using the same degrees of freedom, we optimize for \( \alpha(\omega) \) to obtain \( Z_1(\omega) \) using (9). While in ANC, \( Y_2(\omega) \) is constrained to be the noise reference, such a choice need not be made a-priori in the proposed method. Further, we use a more flexible minimum tracking approach to determine noise-only segments. We now obtain an expression for \( E[\hat{Z}_1(\omega) \hat{Z}_1^*(\omega)] \). We have

\[
\eta(\omega) = E[\hat{Z}_1(\omega) \hat{Z}_1^*(\omega)]
\]

\[
= E[(\alpha(\omega) Y_1(\omega) + (1 - \alpha(\omega)) Y_2(\omega))]^*
\]

\[
= \alpha(\omega) R^{11}_Y(\omega) \alpha^*(\omega) + \alpha(\omega) R^{12}_Y(\omega)(1 - \alpha(\omega))^*
\]

\[
+ (1 - \alpha(\omega)) R^{21}_Y(\omega) \alpha^*(\omega)
\]

\[
+ (1 - \alpha(\omega)) R^{22}_Y(\omega)(1 - \alpha(\omega))^*,
\]

where \( R^{ij}_Y(\omega) \) corresponds to the \( (i, j)^{th} \) element of the matrix \( R_Y(\omega) = E[\hat{Y}(\omega) \hat{Y}^*(\omega)] \). The matrix \( R_Y(\omega) \) is the cross spectral density of the observed microphone signals during regions of speech absence, i.e., it is the noise cross spectral density. Differentiating \( \eta(\omega) \) with respect to \( \alpha(\omega) \) (the conjugate gradient [7, App. B] and equating the result to zero provides

\[
\hat{\alpha}(\omega) = \frac{R^{11}_Y(\omega) - R^{12}_Y(\omega)}{R^{11}_Y(\omega) + R^{22}_Y(\omega) - R^{12}_Y(\omega) - R^{21}_Y(\omega)}.
\]

\[
\hat{\alpha}(\omega) = \frac{R^{11}_Y(\omega) - R^{12}_Y(\omega)}{R^{11}_Y(\omega) + R^{22}_Y(\omega) - R^{12}_Y(\omega) - R^{21}_Y(\omega)}.
\]
An estimate of the clean speech signal is obtained from the microphone signals $Y_1(\omega)$ and $Y_2(\omega)$ as

$$\hat{S}(\omega) = \tilde{Z}_i(\omega) - \hat{a}(\omega)(Y_1(\omega) + (1 - \hat{a}(\omega))Y_2(\omega)).$$  \hspace{1cm} (13)

We observe that $\hat{a}(\omega)$ has been derived such that the noise energy in $Z_i(\omega)$ is minimized. To obtain $\hat{a}(\omega)$ using (12), we need an estimate of the noise cross spectral density matrix, which is addressed in section 3.2.

Once $\hat{a}(\omega)$ has been obtained, the speech signal is estimated in a single step using (13). This is in contrast to most BSS algorithms that employ some form of gradient descent and thus require several iterations to converge.

### 3.2. Estimating the noise spectral density

If we let $R_{ij}(\omega)$ denote the $(i,j)^{th}$ element of the matrix $R_Y(\omega) = E[Y(\omega)Y^*(\omega)]$, we have from (6)

$$R_{11}(\omega) = |a_{11}(\omega)|^2 P_S(\omega) + |a_{12}(\omega)|^2 P_N(\omega)$$

$$R_{12}(\omega) = a_{11}(\omega)a_{21}(\omega)P_S(\omega) + a_{12}(\omega)a_{22}(\omega)P_N(\omega)$$

$$R_{21}(\omega) = a_{21}(\omega)a_{11}(\omega)P_S(\omega) + a_{22}(\omega)a_{12}(\omega)P_N(\omega)$$

$$R_{22}(\omega) = |a_{21}(\omega)|^2 P_S(\omega) + |a_{22}(\omega)|^2 P_N(\omega),$$  \hspace{1cm} (14)

where $P_S(\omega) = E[S(\omega)S^*(\omega)]$ and $P_N(\omega) = E[N(\omega)N^*(\omega)]$ are the power spectral densities of the speech and noise signals respectively. The cross terms vanish due to the independence assumption, i.e., $E[S(\omega)N^*(\omega)] = E[S^*(\omega)N(\omega)] = 0$.

We note that $R_{11}(\omega)$ and $R_{22}(\omega)$ are both real quantities that attain their minimum values when the speech signal is absent. Thus by tracking the minimum of either $R_{11}(\omega)$ or $R_{22}(\omega)$, noise only segments can be identified. From these segments $R_{11}(\omega)$ and $R_{22}(\omega)$ can be estimated and $\hat{a}(\omega)$ can be obtained using (12). $R_{11}(\omega)$ can thus be used in an approach similar to the the minimum statistics algorithm [8]. A buffer of D past cross spectral densities is maintained for each frequency bin and the minimum is tracked in this buffer. The buffer size D should be large enough to include non-speech regions and small enough to account for nonstationary noise. Such an approach has been used in [9] to compute the noise cross spectral density for bias compensation in blind source separation of multiple speakers in an additive noise environment.

As an alternative to minimum tracking, a reliable voice activity detector (VAD) can be employed to identify noise-only regions. However, a VAD imposes a stronger condition than is required to update the noise spectral density in each bin; a VAD requires speech energy to be absent in all frequency bins simultaneously to update the noise estimate. The minimum tracking approach on the other hand exploits the fact that speech energy is not present in all frequency bins at all times. Thus by tracking the minimum in each frequency bin, the estimates can be updated during periods of binwise speech inactivity.

### 3.3. Interpreting $\hat{a}(\omega)$

In section 3.1, we derived $\hat{a}(\omega)$ such that the noise energy was minimized in the output signal. Here we analyze the resulting signal in more detail by considering a few special cases.

#### 3.3.1. Case 1: $a_{ij}(\omega) \neq 0 \quad i,j = 1,2$

This is the most typical case, where noisy speech is present in both channels $Y_1(\omega)$ and $Y_2(\omega)$. Setting $P_S(\omega) = 0$ in (14), we obtain the noise cross spectral density matrix

$$R_Y(\omega) = \begin{bmatrix} |a_{11}(\omega)|^2 P_S(\omega) & a_{12}(\omega)a_{21}(\omega)P_N(\omega) \\ a_{21}(\omega)a_{12}(\omega)P_N(\omega) & |a_{22}(\omega)|^2 P_N(\omega) \end{bmatrix},$$  \hspace{1cm} (15)

From (15) and (12) and after some simple algebra, we obtain

$$\hat{a}(\omega) = \frac{a_{22}(\omega)}{a_{22}(\omega) - a_{12}(\omega)}. \hspace{1cm} (16)$$

Applying (13) to the microphone signals $Y_1(\omega)$ and $Y_2(\omega)$, we get after simplification

$$\hat{S}(\omega) = \frac{a_{11}(\omega)a_{22}(\omega) - a_{12}(\omega)a_{21}(\omega)}{a_{22}(\omega) - a_{12}(\omega)} S(\omega), \hspace{1cm} (17)$$

which corresponds to a filtered version (scaled in the instantaneous mixing case) of the original speech signal. The contribution of the noise source is cancelled. As is well known, in general, BSS methods for convolutive mixtures can recover the sources only up to a filtering.

#### 3.3.2. Case 2: $a_{12}(\omega) = a_{21}(\omega) = 0$

In this case, we have $Y_1(\omega) = a_{11}(\omega)S(\omega)$ and $Y_2(\omega) = a_{22}(\omega)N(\omega)$, i.e., the speech and noise signals are already separated. From (14), by setting $a_{12}(\omega) = a_{21}(\omega) = 0$ and $P_S(\omega) = 0$, we can write the noise cross spectral density matrix as

$$R_Y(\omega) = \begin{bmatrix} 0 & 0 \\ 0 & |a_{22}(\omega)|^2 P_N(\omega) \end{bmatrix}, \hspace{1cm} (18)$$

so that from (12) we have $\hat{a}(\omega) = 1$. Thus from (13), we have

$$\hat{S}(\omega) = \hat{a}(\omega)Y_1(\omega) = a_{11}(\omega)S(\omega), \hspace{1cm} (19)$$

and the speech signal at the microphone is recovered without any modification.

If the position of the speech and noise signals at the microphones is reversed, we have $Y_1(\omega) = a_{11}(\omega)N(\omega)$ and $Y_2(\omega) = a_{22}(\omega)S(\omega)$. In this case,

$$R_Y(\omega) = \begin{bmatrix} |a_{11}(\omega)|^2 P_N(\omega) & 0 \\ 0 & 0 \end{bmatrix}, \hspace{1cm} (20)$$

and thus $\hat{a}(\omega) = 0$ so that $\hat{S}(\omega) = (1 - \hat{a}(\omega))Y_2(\omega) = a_{22}(\omega)S(\omega)$, as expected.

#### 3.3.3. Case 3: $N(\omega) = 0$

This corresponds to the case when there is no background noise. The noise cross spectral density matrix is zero, which results in unreliable estimates of $\hat{a}(\omega)$. This problem can be solved by regularizing $R_Y(\omega)$ as $R_Y^{\alpha}(\omega) = R_Y(\omega) + \epsilon I$, where I is the $2 \times 2$ identity matrix and $0 < \epsilon \ll 1$. The resulting estimate is $\hat{a}(\omega) = 0.5$ and $\hat{S}(\omega) = 0.5(a_{11}(\omega)S(\omega) + a_{12}(\omega)S(\omega))$.

### 3.4. Permutation problem

In the method proposed in this paper, we exploit the knowledge that the input consists of two specific signals, speech and noise. Unlike other methods that attempt to recover both signals, we estimate only the speech signal using equation (13). Consequently, this approach does not suffer from the permutation problem that is common to typical convolutive BSS methods operating in the frequency domain.
4. Experiments

To evaluate the proposed noise reduction approach, experiments were performed using four speech utterances, two male and two female, each approximately 10s long. The speech utterances were sampled at 8kHz and artificially mixed with white noise. Prior to the mixing, the speech and noise signals were normalized to have a signal-to-noise ratio of 0 dB. The convolutive mixing was performed using finite impulse response (FIR) filters of orders five and ten. The filter coefficients were generated from an exponentially decaying Cauchy noise generator.

We also used room impulse responses generated artificially using the source-image model as in [10]. The dimension of the room was $10 \times 10 \times 10$ meters. The sources were positioned at $[1,5, 1, 1.5]$ and $[2.5, 1, 1.5]$ and the microphones at $[1.5, 3, 1.5]$ and $[2.5, 3, 1.5]$. The cross spectral density matrix $R_Y(\omega)$ was evaluated using sample averages as in [6] on a frame-by-frame basis. A frame length of 512 samples (1024 samples for the simulated room) was used with a 50% overlap. For each frame, $R_Y(\omega)$ was computed by averaging over five neighboring frames. In the current work, an offline implementation was considered and $\delta(\omega)$ was estimated once for the entire signal. The frames were Hann windowed prior to the computation of $R_Y(\omega)$. $Y_1(\omega)$ and $Y_2(\omega)$ were also computed on a frame-by-frame basis by applying the discrete Fourier transform to Hann windowed segments of $y_1$ and $y_2$.

The separated speech signal was estimated for each frame according to (13). After the separation, the inverse discrete Fourier transform was applied to $S(\omega)$ and the enhanced time domain signal was obtained through the overlap-add procedure.

The performance of the proposed method was evaluated using the signal-to-interference ratio (SIR). The SIR at the $i^{th}$ output channel is defined as [6]

$$SIR_i = 10 \log_{10} \frac{\sum_{t=1}^{T} \hat{z}_{i1}(t)^2}{\sum_{t=1}^{T} \hat{z}_{i2}(t)^2}$$

where $\hat{z}_{i1}(t)$ is the output at the $i^{th}$ channel (after mixing and demixing) when only the speech signal is active and $\hat{z}_{i2}(t)$ is the output when only the noise signal is active, and $T$ is the number of samples in the utterance. In the proposed method, there is only one output channel since we only estimate the speech.

<table>
<thead>
<tr>
<th>Filter length</th>
<th>Min. tracking</th>
<th>Ref. method [6]</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>14.9 ± 0.9</td>
<td>8.6 ± 1.7</td>
</tr>
<tr>
<td>10</td>
<td>15.6 ± 0.8</td>
<td>8.8 ± 1.9</td>
</tr>
<tr>
<td>Room</td>
<td>11.1 ± 1.1</td>
<td>8.9 ± 0.9</td>
</tr>
</tbody>
</table>

Table 1: Mean SIR improvement and 95% confidence intervals, averaged over ten realizations of the 5 and 10 tap mixing filters, and for the simulated room impulse responses (approx. 1000 taps at 8kHz).

Table 1 shows the improvement in SIR due to the separation. For comparisons, we also include results obtained from the method of [6]. The frame length and length of unmixing filters were (1024,128) for the short mixing filters and (2048,1024) for the simulated room. It can be seen that there is a significant advantage due to the proposed scheme. As seen in Fig. 2, which shows the improvement in SIR for a typical realization with a 10 tap filter, the proposed method achieves good performance in a single step, without the need for any iterations. Informal listening confirms the improved separation due to our method.

5. Conclusions

In this paper, we have presented a noise reduction technique that is based on BSS. Noisy speech is viewed as a convolutive mixture of speech and noise signals. The demixing coefficients necessary to recover the speech signal are obtained from the noise cross spectral density matrix, which is estimated through minimum tracking. By estimating only the speech signal, the permutation problem is avoided. Separation is achieved in a single step leading to significant computational savings compared to most BSS algorithms, which require several iterations. Experiments show significant gains in SIR compared to a state-of-the-art BSS technique. Future work will address the arbitrary filtering aspect of convolutive BSS using the minimal distortion principle [11].

6. References