Covariance smoothing and consistent Wiener filtering for artifact reduction in audio source separation

Emmanuel Vincent

METISS Team
Inria Rennes - Bretagne Atlantique
1 Artifacts

2 Covariance smoothing

3 Consistent Wiener filtering

4 Conclusion
Under-determined source separation

We aim to separate $J$ sources from $I < J$ mixture channels.

In the time-frequency domain,

$$x_{nf} = \sum_{j=1}^{J} s_{jnf}$$

- $n$: time frame index
- $f$: frequency bin index
- $x_{nf}$: $I \times 1$ mixture STFT coeffs.
- $s_{jnf}$: $I \times 1$ spatial image of source $j$

Separation is typically performed in two steps:

1. estimate the parameter values of some parametric source model,
2. derive the sources in each time-frequency bin by
   - binary/soft masking,
   - local mixing inversion restricted to a subset of sources,
   - or Wiener filtering.
Artifacts

Separate masking/filtering of each time-frequency bin results in phase and amplitude discontinuities between neighboring bins perceived as artifacts.

Mixture

Binary masking

These artifacts are very annoying for, e.g.,
- hearing-aid speech processing,
- high-fidelity music processing.

Fewer artifacts are often preferred at the expense of increased interference.
General approaches to artifact reduction

Two complementary approaches to artifact reduction:

- smooth the parameter values of the source models,
- given the parameter values of the source models, smooth the masks/filters.
Artifact reduction in the Gaussian modeling framework

We focus on the Gaussian modeling framework (includes NMF, GMM, beamforming, some ICA, etc).

Each source is parameterized by a time- and frequency-dependent multichannel covariance matrix $R_{s_{jnf}}$:

$$s_{jnf} \sim \mathcal{N}(s_{jnf}; 0, R_{s_{jnf}}).$$

Given estimates $\hat{R}_{s_{jnf}}$ of $R_{s_{jnf}}$, we replace the conventional Wiener filter

$$\hat{W}_{jnf} = \hat{R}_{s_{jnf}} \hat{R}_{x_{jnf}}^{-1} \quad \text{with} \quad \hat{R}_{x_{jnf}} = \sum_{j=1}^{J} \hat{R}_{s_{jnf}}$$

by a smoothed Wiener filter $\tilde{W}_{jnf}$.

We then recover the sources by

$$\tilde{s}_{jnf} = \tilde{W}_{jnf} x_{jnf}.$$
Covariance smoothing
Two families of smoothing techniques have been developed in the field of speech enhancement:

- multichannel spatial smoothing with spatially diffuse background,
- single-channel temporal smoothing with stationary background.

In the following, we

- extend temporal smoothing techniques to multi-channel mixtures,
- evaluate both families of techniques on under-determined mixtures involving directional nonstationary interference.
Spatial filter smoothing

Idea: smooth the spatial response by interpolating between the conventional Wiener filter and the identity filter.

\[
\hat{W}^{\text{SFS}}_{\text{jnf}} = (1 - \mu)\hat{W}_{\text{jnf}} + \mu I.
\]

This is equivalent to adding part of the mixture signal back to the estimated source signals.

Chen, J., Benesty, J., Huang, Y., Doclo, S.
New insights into the noise reduction Wiener filter
*IEEE TASLP*, 2006

Rickard, S.T.
The DUET blind source separation algorithm
*Blind Speech Separation*, 2007
Spatial covariance smoothing

Same idea, but nonlinear interpolation.

\[
\widehat{W}_{jnf}^{SCS} = \hat{R}_{s_{jnf}} \left[ (1 - \mu)\hat{R}_{x_{nf}} + \mu \hat{R}_{s_{jnf}} \right]^{-1}.
\]

Doclo, S., Moonen, M.
On the output SNR of the speech-distortion weighted multichannel Wiener filter
*IEEE SPL*, 2005
Temporal filter smoothing

Idea: smooth the power response by interpolating between the conventional Wiener filters for neighboring time frames.

\[
\hat{W}_{TFS}^{jnf} = \frac{1}{L+1} \sum_{l=-L/2}^{L/2} \hat{W}_{j,n+l,f}.
\]

Proc. ICSLP, 2002

Multichannel generalization by replacing single-channel Wiener filters by multichannel Wiener filters.
Temporal covariance smoothing

Same idea, but applied to the source covariance matrices.

\[
\tilde{R}_{s_{jnf}} = \frac{1}{L + 1} \sum_{l=-L/2}^{L/2} \hat{R}_{s_{j,n+l,f}}
\]

\[
\tilde{W}_{jnf}^{TCS} = \tilde{R}_{s_{jnf}} \tilde{R}_{x_{nf}}^{-1} \quad \text{with} \quad \tilde{R}_{x_{nf}} = \sum_{j=1}^{J} \tilde{R}_{s_{jnf}}
\]

Yu, G., Mallat, S., Bacry, E.
Audio denoising by time-frequency block thresholding
*IEEE TSP*, 2008

Multichannel generalization by replacing scalar variances by multichannel covariances.
Temporal SNR smoothing

Same idea, but applied to the Signal-to-Noise Ratio (SNR).

\[
\hat{G}_{jnf} = \hat{R}_{s_{jnf}} [\hat{R}_{x_{nf}} - \hat{R}_{s_{jnf}}]^{-1}
\]

\[
\tilde{G}_{jnf} = \frac{1}{L+1} \sum_{l=-L/2}^{L/2} \hat{G}_{j,n+l,f}
\]

\[
\tilde{W}_{jnf}^{\text{TRS}} = I - [\tilde{G}_{jnf} + I]^{-1}
\]

Ephraim, Y., Malah, D.
Speech enhancement using a minimum mean square error short-time spectral amplitude estimator

*IEEE TASSP*, 1984

Multichannel generalization by replacing scalar SNRs by ”multichannel SNRs” \( G_{jnf} \).
Experimental evaluation (1)

Data: four instantaneous stereo ($I = 2$) mixtures of $J = 3$ sources with known mixing matrix.

Algorithms:
- initial estimation by binary masking or $\ell_1$-norm minimization,
- reestimation by spatial or temporal smoothing.

Metrics:
- Signal-to-Artifacts Ratio (SAR),
- Signal-to-Interference Ratio (SIR).
Experimental evaluation (2)

Binary masking

\[ \text{SAR (dB)} \]
\[ \text{SIR (dB)} \]

\[ \text{SIR (dB)} \]
\[ \text{SAR (dB)} \]

\[ \text{SIR (dB)} \]
\[ \text{SAR (dB)} \]

**Legend:**
- red: spatial filter smoothing
- blue circle: spatial covariance smoothing
- red dashed: temporal filter smoothing
- blue dashed: temporal covariance smoothing
- green dashed: temporal SNR smoothing
Consistent Wiener filtering
STFT consistency

So far, we have treated each time-frequency bin independently. But the STFT is a **redundant** representation!

The sources \( \hat{s}_j \) estimated by conventional Wiener filtering belong to

\[
\mathcal{S} = \{ s \in \mathbb{C}^{I \times N \times F} \text{ s.t. } s_{n,F-f} = \bar{s}_{nf} \quad \forall n, f \}.
\]

Can we smooth the filter so as to obtain a consistent estimate \( \tilde{s}_j \), i.e., \( \tilde{s}_j \in \text{STFT}(\mathbb{R}^{I \times T}) \subsetneq \mathcal{S} \) is the STFT of a time-domain signal?

\[
s \text{ is consistent } \iff \mathcal{F}(s) = 0
\]

where \( \mathcal{F} : s \mapsto s - \text{STFT}(i\text{STFT}(s)) \) is a projector.
Consistent Wiener filtering

General formulation (1)

To do so, let us come back to the definition of the Wiener filter.

Let \( S_{nf} = \begin{bmatrix} s_{1,nf} \\ \vdots \\ s_{J-1,nf} \end{bmatrix} \).

The Wiener filter computes the mode of \( P(S_{nf} | x_{nf}) \) assuming that \( s_{jnf} \) are independent zero-mean Gaussian with estimated covariance \( \hat{R}_{s_{jnf}} \).

This distribution is Gaussian with mean \( \hat{\mu}_{nf} \) and precision \( \Lambda_{nf} \) given by

\[
\begin{align*}
\hat{\mu}_{nf} &= \Sigma_{Sx} \Sigma_{xx}^{-1} x_{nf}, \\
\Lambda_{nf} &= (\Sigma_{SS} - \Sigma_{Sx} \Sigma_{xx}^{-1} \Sigma_{xS})^{-1}
\end{align*}
\]

\[
\begin{align*}
\Sigma_{xS} &= \Sigma_{Sx}^H = \begin{bmatrix} \hat{R}_{s_{1,nf}} & \cdots & \hat{R}_{s_{J-1,nf}} \end{bmatrix} \\
\Sigma_{xx} &= \sum_{j=1}^{J} \hat{R}_{s_{jnf}} \\
\Sigma_{SS} &= \text{diag}(\hat{R}_{s_{1,nf}}, \cdots, \hat{R}_{s_{J-1,nf}})
\end{align*}
\]
**General formulation (2)**

Wiener filtering is thus equivalent to minimizing $-\log P(S|x)$, which is equal up to a constant to the **quadratic loss**

$$\psi(S) = \sum_{n,f} (S_{nf} - \hat{\mu}_{nf})^H \Lambda_{nf} (S_{nf} - \hat{\mu}_{nf}).$$

Without the consistency constraint, the solution is obviously

$$\arg\min_{S \in S} \psi(S) = \hat{\mu}.$$  

We aim to solve one of the following constrained problems instead:

$$\arg\min_{S \in S} \psi(S) \quad \text{s.t.} \quad \mathcal{F}(S) = 0 \quad \text{(hard constraint)}$$

$$\mathcal{F}(S) \text{ "small" (soft penalty)}$$
**Consistency as a hard constraint**

We formulate the hard constrained problem in the time domain as

$$\arg \min_{\hat{S} \in \mathbb{R}^{J \times I \times T}} \psi(\text{STFT}(\hat{S}))$$

The solution $\tilde{\hat{S}}$ satisfies $\text{STFT}^* \circ \Lambda \circ \text{STFT}(\tilde{\hat{S}}) = \text{STFT}^* \circ \Lambda(\hat{\mu})$ where $\text{STFT}^*$ is the adjoint of the STFT and $\Lambda(S)_{nf} = \Lambda_{nf} S_{nf}$.

When the STFT analysis and synthesis windows are identical, $\text{STFT}^*$ is equal to $\text{iSTFT}$ up to a constant, hence

$$\text{iSTFT} \circ \Lambda \circ \text{STFT}(\tilde{\hat{S}}) = \text{iSTFT} \circ \Lambda(\hat{\mu})$$

We invert $\text{iSTFT} \circ \Lambda \circ \text{STFT}$ using the conjugate gradient (CG) algorithm with preconditioning

$$M^{-1} : \hat{S} \mapsto \text{iSTFT} \circ \Lambda^{-1} \circ \text{STFT}(\hat{S}).$$
Consistent Wiener filtering

Consistency as a soft penalty

Idea: avoid hard constraints when the source parameters are unknown.

Consider the following time-frequency domain penalized problem instead:

$$\arg\min_{S \in S} \psi(S) + \gamma \|F(S)\|^2$$

The solution $\tilde{S}$ satisfies $(\Lambda + \gamma F^* \circ F)(\tilde{S}) = \Lambda(\hat{\mu})$.

When the STFT analysis and synthesis windows are identical, $F$ is Hermitian, hence $F^* \circ F = F$ and

$$(\Lambda + \gamma F)(\tilde{S}) = \Lambda(\hat{\mu})$$

We invert $\Lambda + \gamma F$ using CG with preconditioning

$$M^{-1}: S \mapsto \left(\Lambda + \gamma \frac{F^T N - T}{F N} \text{Id}\right)^{-1}(S).$$
Experimental evaluation (1)

Data: mono ($l = 1$) mixtures of speech and noise at 3 different SNRs.

STFT: half-overlapping sine windows of length 1024 (64 ms).

Source covariance estimates:
- oracle
- or blind (spectral subtraction).

Algorithm: 7 different values of $\gamma$, hard constrained version formally denoted as $\gamma = +\infty$. 

Experimental results (2)

Condition numbers with \((+P)\) and without \((-P)\) preconditioning

![Graph showing condition numbers with \((+P)\) and without \((-P)\) preconditioning.](image-url)
Experimental results (3)

Signal-to-Distortion Ratio (SDR)

- Oracle +10 dB
- Oracle 0 dB
- Oracle -10 dB
- Blind +10 dB
- Blind 0 dB
- Blind -10 dB
### Experimental results (4)

<table>
<thead>
<tr>
<th>Input SNR</th>
<th>-10 dB</th>
<th>0 dB</th>
<th>+10 dB</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
<td>Time</td>
</tr>
<tr>
<td>Wiener</td>
<td>11.5</td>
<td>21.2</td>
<td>12.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Oracle</td>
<td>12.4</td>
<td>24.0</td>
<td>12.9</td>
<td>1.7</td>
</tr>
<tr>
<td>Oracle</td>
<td>12.1</td>
<td>24.2</td>
<td>12.5</td>
<td>2.4</td>
</tr>
<tr>
<td>Oracle</td>
<td>11.7</td>
<td>22.6</td>
<td>12.3</td>
<td>2.5</td>
</tr>
<tr>
<td>Oracle</td>
<td>12.7</td>
<td>25.4</td>
<td>13.6</td>
<td>2.7</td>
</tr>
<tr>
<td>Oracle</td>
<td>5.0</td>
<td>6.1</td>
<td>11.6</td>
<td>0.1</td>
</tr>
<tr>
<td>Oracle</td>
<td>4.9</td>
<td>6.3</td>
<td>10.7</td>
<td>1.9</td>
</tr>
<tr>
<td>Oracle</td>
<td>4.9</td>
<td>6.9</td>
<td>9.5</td>
<td>6.6</td>
</tr>
<tr>
<td>Oracle</td>
<td>6.6</td>
<td>10.0</td>
<td>9.3</td>
<td>7.8</td>
</tr>
<tr>
<td>Oracle</td>
<td>8.3</td>
<td>13.9</td>
<td>9.7</td>
<td>6.4</td>
</tr>
<tr>
<td>Oracle</td>
<td>14.7</td>
<td>16.0</td>
<td>20.8</td>
<td>0.1</td>
</tr>
<tr>
<td>Oracle</td>
<td>14.7</td>
<td>16.2</td>
<td>19.9</td>
<td>1.3</td>
</tr>
<tr>
<td>Oracle</td>
<td>14.7</td>
<td>16.5</td>
<td>19.4</td>
<td>2.0</td>
</tr>
<tr>
<td>Oracle</td>
<td>15.8</td>
<td>18.7</td>
<td>19.1</td>
<td>2.5</td>
</tr>
<tr>
<td>Oracle</td>
<td>17.3</td>
<td>21.7</td>
<td>19.4</td>
<td>2.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SIR</th>
<th>SAR</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiener</td>
<td>-4.8</td>
<td>5.9</td>
<td>0.1</td>
</tr>
<tr>
<td>MISI</td>
<td>-4.8</td>
<td>4.9</td>
<td>2.1</td>
</tr>
<tr>
<td>PPR</td>
<td>-4.9</td>
<td>2.8</td>
<td>11.2</td>
</tr>
<tr>
<td>ISSIR</td>
<td>-2.3</td>
<td>1.0</td>
<td>17.8</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.5</td>
<td>3.3</td>
<td>1.2</td>
</tr>
<tr>
<td>Proposed</td>
<td>4.9</td>
<td>9.6</td>
<td>11.2</td>
</tr>
<tr>
<td>Proposed</td>
<td>6.6</td>
<td>10.0</td>
<td>9.3</td>
</tr>
<tr>
<td>Proposed</td>
<td>8.3</td>
<td>13.9</td>
<td>9.7</td>
</tr>
<tr>
<td>Proposed</td>
<td>14.7</td>
<td>16.0</td>
<td>20.8</td>
</tr>
<tr>
<td>Proposed</td>
<td>14.7</td>
<td>16.2</td>
<td>19.9</td>
</tr>
<tr>
<td>Proposed</td>
<td>14.7</td>
<td>16.5</td>
<td>19.4</td>
</tr>
<tr>
<td>Proposed</td>
<td>15.8</td>
<td>18.7</td>
<td>19.1</td>
</tr>
<tr>
<td>Proposed</td>
<td>17.3</td>
<td>21.7</td>
<td>19.4</td>
</tr>
</tbody>
</table>

Gunawan, D., Sen, D.

Iterative phase estimation for the synthesis of separated sources from single-channel mixtures

*IEEE SPL*, 2010

Sturmel, N., Daudet, L.

Iterative phase reconstruction of Wiener filtered signals in *Proc. ICASSP*, 2012.

Sturmel, N., Daudet, L.

Informed source separation using iterative reconstruction

arXiv:1202.2075v1
Conclusion
Conclusion

Among the tested smoothing techniques, temporal covariance smoothing provides the best tradeoff between artifacts and interference independently of the initial separation algorithm.

Consistent Wiener filtering can also decrease artifacts and interference.

Both techniques rely on one parameter (kernel width $L$ or tradeoff $\gamma$) whose setting heavily depends on the accuracy of the initial source covariance estimates $\hat{R}_{s_{inf}}$. 