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,

$$\mathbf{x}_{nf} = \sum_{j=1}^{J} \mathbf{s}_{jnf}$$

n: time frame index
f: frequency bin index
$\mathbf{x}_{nf}$: $I \times 1$ mixture STFT coeffs.
$\mathbf{s}_{jnf}$: $I \times 1$ 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 $\mathbf{R}_{\mathbf{s}_{jnf}}$:

$$\mathbf{s}_{jnf} \sim \mathcal{N}(\mathbf{s}_{jnf}; \mathbf{0}, \mathbf{R}_{\mathbf{s}_{jnf}}).$$

Given estimates $\widehat{\mathbf{R}}_{\mathbf{s}_{jnf}}$ of $\mathbf{R}_{\mathbf{s}_{jnf}}$, we replace the conventional Wiener filter

$$\widehat{\mathbf{W}}_{jnf} = \widehat{\mathbf{R}}_{\mathbf{s}_{jnf}} \widehat{\mathbf{R}}_{\mathbf{x}_{nf}}^{-1} \quad \text{with} \quad \widehat{\mathbf{R}}_{\mathbf{x}_{nf}} = \sum_{j=1}^{J} \widehat{\mathbf{R}}_{\mathbf{s}_{jnf}}$$

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

We then recover the sources by

$$\widetilde{\mathbf{s}}_{jnf} = \widetilde{\mathbf{W}}_{jnf} \mathbf{x}_{nf}.$$

Covariance smoothing

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.

$$\widetilde{\mathbf{W}}_{jnf}^{\mathrm{SFS}} = (1-\mu)\widehat{\mathbf{W}}_{jnf} + \mu\,\mathbf{I}.$$

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.

$$\widetilde{\mathbf{W}}_{jnf}^{\mathrm{SCS}} = \widehat{\mathbf{R}}_{\mathbf{s}_{jnf}}[(1-\mu)\widehat{\mathbf{R}}_{\mathbf{x}_{nf}} + \mu\widehat{\mathbf{R}}_{\mathbf{s}_{jnf}}]^{-1}.$$

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.

$$\widetilde{\mathbf{W}}_{jnf}^{\mathrm{TFS}} = \frac{1}{L+1} \sum_{l=-L/2}^{L/2} \widehat{\mathbf{W}}_{j,n+l,f}.$$

Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., et al.
Qualcomm-ICSI-OGI features for ASR
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.

$$\widetilde{\mathbf{R}}_{\mathbf{s}_{jnf}} = \frac{1}{L+1} \sum_{l=-L/2}^{L/2} \widehat{\mathbf{R}}_{\mathbf{s}_{j,n+l,f}}$$

$$\widetilde{\mathbf{W}}_{jnf}^{\mathrm{TCS}} = \widetilde{\mathbf{R}}_{\mathbf{s}_{jnf}} \widetilde{\mathbf{R}}_{\mathbf{x}_{nf}}^{-1} \quad \text{with} \quad \widetilde{\mathbf{R}}_{\mathbf{x}_{nf}} = \sum_{j=1}^{J} \widetilde{\mathbf{R}}_{\mathbf{s}_{jnf}}$$

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).

$$\widehat{\mathbf{G}}_{jnf} = \widehat{\mathbf{R}}_{\mathbf{s}_{jnf}}[\widehat{\mathbf{R}}_{\mathbf{x}_{nf}} - \widehat{\mathbf{R}}_{\mathbf{s}_{jnf}}]^{-1}$$

$$\widetilde{\mathbf{G}}_{jnf} = \frac{1}{L+1} \sum_{l=-L/2}^{L/2} \widehat{\mathbf{G}}_{j,n+l,f}$$

$$\widetilde{\mathbf{W}}_{jnf}^{\mathrm{TRS}} = \mathbf{I} - [\widetilde{\mathbf{G}}_{jnf} + \mathbf{I}]^{-1}$$

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" $\mathbf{G}_{jnf}$.

Experimental evaluation (1)

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

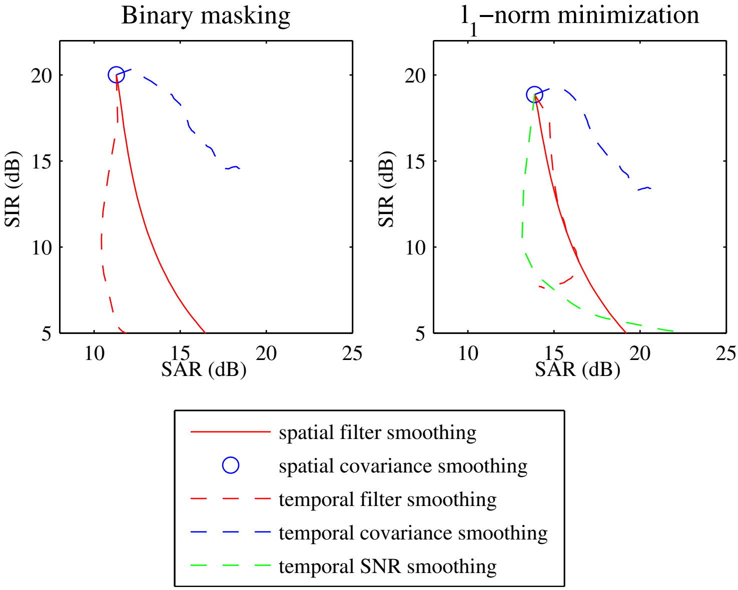
Algorithms:

- initial estimation by binary masking or ℓ_1-norm minimization,
- reestimation by spatial or temporal smoothing.

Metrics:

- Signal-to-Artifacts Ratio (SAR),
- Signal-to-Interference Ratio (SIR).

Experimental evaluation (2)



Consistent Wiener filtering

STFT consistency

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

The sources $\widehat{\mathbf{s}}_j$ estimated by conventional Wiener filtering belong to

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

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

$$\mathbf{s} \text{ is consistent} \quad \Leftrightarrow \quad \mathcal{F}(\mathbf{s}) = 0$$

where $\mathcal{F} : \mathbf{s} \mapsto \mathbf{s} - \mathrm{STFT}(\mathrm{iSTFT}(\mathbf{s}))$ is a projector.

General formulation (1)

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

Let $\mathbf{S}_{nf} = \begin{bmatrix} \mathbf{s}_{1,nf} \\ \vdots \\ \mathbf{s}_{J-1,nf} \end{bmatrix}$.

The Wiener filter computes the mode of $P(\mathbf{S}_{nf}|\mathbf{x}_{nf})$ assuming that $\mathbf{s}_{jnf}$ are independent zero-mean Gaussian with estimated covariance $\widehat{\mathbf{R}}_{\mathbf{s}_{jnf}}$.

This distribution is Gaussian with mean $\hat{\boldsymbol{\mu}}_{nf}$ and precision $\boldsymbol{\Lambda}_{nf}$ given by

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{nf} &= \boldsymbol{\Sigma}_{\mathbf{Sx}} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \mathbf{x}_{nf}, \\ \boldsymbol{\Lambda}_{nf} &= (\boldsymbol{\Sigma}_{\mathbf{SS}} - \boldsymbol{\Sigma}_{\mathbf{Sx}} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\Sigma}_{\mathbf{xS}})^{-1} \end{aligned} \quad \text{with} \quad \begin{aligned} \boldsymbol{\Sigma}_{\mathbf{xS}} &= \boldsymbol{\Sigma}_{\mathbf{Sx}}^{H} = \left[\widehat{\mathbf{R}}_{\mathbf{s}_{1,nf}}, \dots, \widehat{\mathbf{R}}_{\mathbf{s}_{J-1,nf}}\right] \\ \boldsymbol{\Sigma}_{\mathbf{xx}} &= \sum_{j=1}^{J} \widehat{\mathbf{R}}_{\mathbf{s}_{jnf}} \\ \boldsymbol{\Sigma}_{\mathbf{SS}} &= \operatorname{diag}(\widehat{\mathbf{R}}_{\mathbf{s}_{1,nf}}, \dots, \widehat{\mathbf{R}}_{\mathbf{s}_{J-1,nf}}) \end{aligned}$$

General formulation (2)

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

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

Without the consistency constraint, the solution is obviously

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

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

$$\arg\min_{\mathbf{S}\in\mathcal{S}} \psi(\mathbf{S}) \quad \text{s.t.} \quad \begin{array}{l} \mathcal{F}(\mathbf{S}) = 0 \text{ (hard constraint)} \\ \mathcal{F}(\mathbf{S}) \text{ ``small'' (soft penalty)} \end{array}$$

Consistency as a hard constraint

We formulate the hard constrained problem in the time domain as

$$\boxed{\arg\min_{\check{\mathbf{S}}\in\mathbb{R}^{J-1\times I\times T}} \psi(\mathrm{STFT}(\check{\mathbf{S}}))}$$

The solution $\widetilde{\check{\mathbf{S}}}$ satisfies $\mathrm{STFT}^* \circ \mathbf{\Lambda} \circ \mathrm{STFT}(\widetilde{\check{\mathbf{S}}}) = \mathrm{STFT}^* \circ \mathbf{\Lambda}(\hat{\mu})$ where STFT^* is the adjoint of the STFT and $\mathbf{\Lambda}(\mathbf{S})_{nf} = \mathbf{\Lambda}_{nf}\mathbf{S}_{nf}$.

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

$$\boxed{\mathrm{iSTFT} \circ \mathbf{\Lambda} \circ \mathrm{STFT}(\widetilde{\check{\mathbf{S}}}) = \mathrm{iSTFT} \circ \mathbf{\Lambda}(\hat{\mu})}$$

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

$$\mathbf{M}^{-1} : \check{\mathbf{S}} \mapsto \mathrm{iSTFT} \circ \mathbf{\Lambda}^{-1} \circ \mathrm{STFT}(\check{\mathbf{S}}).$$

Consistency as a soft penalty

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

Consider the following time-frequency domain penalized problem instead:

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

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

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

$$\boxed{(\mathbf{\Lambda} + \gamma\mathcal{F})(\widetilde{\mathbf{S}}) = \mathbf{\Lambda}(\hat{\boldsymbol{\mu}})}$$

We invert $\mathbf{\Lambda} + \gamma\mathcal{F}$ using CG with preconditioning

$$\mathbf{M}^{-1} : \mathbf{S} \mapsto \left(\mathbf{\Lambda} + \gamma \frac{FN - T}{FN} \mathrm{Id}\right)^{-1} (\mathbf{S}).$$

Experimental evaluation (1)

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

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

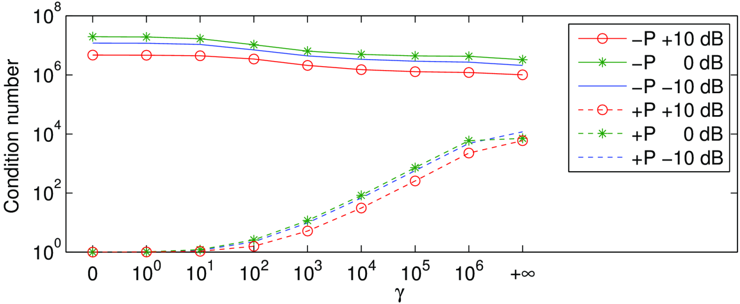
Source covariance estimates:

- oracle
- or blind (spectral subtraction).

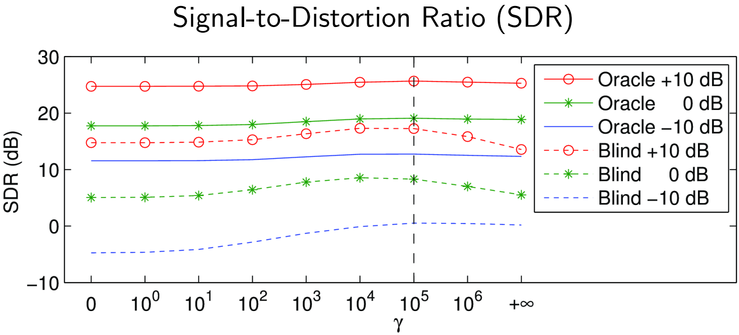
Algorithm: 7 different values of γ, hard constrained version formally denoted as $\gamma = +\infty$.

Experimental results (2)

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



Experimental results (3)



Experimental results (4)

Input SNR		-10 dB				0 dB				+10 dB			
		SDR	SIR	SAR	Time	SDR	SIR	SAR	Time	SDR	SIR	SAR	Time
Oracle	Wiener	11.5	21.2	12.2	0.1	17.7	25.0	18.8	0.1	24.7	30.1	26.4	0.1
	MISI	12.4	24.0	12.9	1.7	18.4	27.2	19.2	1.1	25.3	31.7	26.6	0.9
	PPR	12.1	24.2	12.5	2.4	18.2	26.9	19.0	1.0	25.0	31.2	26.4	0.6
	ISSIR	11.7	22.6	12.3	2.5	17.9	25.7	18.8	1.3	24.8	30.4	26.4	0.7
	Proposed	**12.7**	25.4	13.6	2.7	**19.1**	27.7	20.1	2.0	**25.7**	31.4	27.2	1.1
Blind	Wiener	-4.8	-4.8	5.9	0.1	5.0	6.1	11.6	0.1	14.7	16.0	20.8	0.1
	MISI	-4.8	-4.6	4.9	2.1	4.9	6.3	10.7	1.9	14.7	16.2	19.9	1.3
	PPR	-4.9	-3.8	2.8	11.2	4.9	6.9	9.5	6.6	14.7	16.5	19.4	2.0
	ISSIR	-2.3	-0.8	1.0	17.8	6.6	10.0	9.3	7.8	15.8	18.7	19.1	2.5
	Proposed	**0.5**	3.3	1.2	10.6	**8.3**	13.9	9.7	6.4	**17.3**	21.7	19.4	2.7

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 γ) whose setting heavily depends on the accuracy of the initial source covariance estimates $\widehat{\mathbf{R}}_{\mathbf{s}_{jnf}}$.

References

E. Vincent, “An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation”, in *Proc. LVA/ICA*, pp. 157-164, 2010.

J. Le Roux and E. Vincent, “Consistent Wiener filtering for audio source separation”, *IEEE Signal Processing Letters*, to appear.