An Experimental Evaluation of Wiener Filter Smoothing Techniques Applied to Under-Determined Audio Source Separation

Emmanuel Vincent

INRIA, Centre de Rennes - Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes Cedex, France
emmanuel.vincent@inria.fr

Abstract. Multichannel under-determined source separation is often carried out in the time-frequency domain by estimating the source coefficients in each time-frequency bin based on some sparsity assumption. Due to the limited amount of data, this estimation is often inaccurate and results in musical noise artifacts. A number of single- and multichannel smoothing techniques have been introduced to reduce such artifacts in the context of speech denoising but have not yet been systematically applied to under-determined source separation. We present some of these techniques, extend them to multichannel input when needed, and compare them on a set of speech and music mixtures. Many techniques initially designed for diffuse and/or stationary interference appear to fail with directional nonstationary interference. Temporal covariance smoothing provides the best tradeoff between artifacts and interference and increases the overall signal-to-distortion ratio by up to 3 dB.

1 Introduction

Source separation is the task of recovering the contribution or *spatial image* $\mathbf{c}_j(t)$ of each source indexed by j, $1 \leq j \leq J$, within a multichannel mixture signal $\mathbf{x}(t)$ with

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t). \tag{1}$$

In this paper, we focus on audio source separation [14] in the under-determined setting when the number of mixture channels I is such that $1 < I < J$. This task is typically addressed in the time-frequency domain via the Short-Term Fourier Transform (STFT). A popular approach consists of exploiting the sparsity of audio sources in this domain so as to estimate the STFT coefficients of the most prominent sources in each time-frequency bin and set the STFT coefficients of the other sources to zero. Depending on the assumed number of active sources from 1 to I and on the chosen estimation criterion, this leads to algorithms such as *binary masking* [11], soft masking [3] or ℓ_1*-norm minimization* [17].

Although these separation algorithms succeed at reducing *interference* from unwanted sources, they generate a significant amount of time- and frequency-localized *artifacts* also known as musical noise [15]. These artifacts are particularly annoying in scenarios such as hearing-aid speech processing or high-fidelity music processing where fewer artifacts are preferred at the expense of increased interference. Adaptive time-frequency representations with maximal sparsity or STFTs with increased frame overlap only moderately reduce artifacts [10,2]. Indeed, the localized nature of artifacts is due to the limited amount of data available for estimation in each time-frequency bin, such that similar mixture STFT coefficients in neighboring time-frequency bins may result in very dissimilar estimated source STFT coefficients. This causes strong discontinuities independently of the chosen representation. Joint processing of several time-frequency bins is needed to further reduce artifacts.

One approach consists of modeling the dependencies between the STFT coefficients of the spatial image of each source via some joint probabilistic prior. For instance, these coefficients may be locally modeled as zero-mean Gaussian vector random variables whose covariance matrices are either constant over neighboring time-frequency bins [8,13] or subject to more advanced spectral models including constraints such as harmonicity [14]. These constraints increase the *smoothness* of the estimated source covariance matrices hence of the estimated source STFT coefficients derived by Wiener filtering. Although they typically reduce both interference and artifacts compared to sparsity-based algorithms, these algorithms still result in a significant level of artifacts [15,12].

In this paper, we explore a complementary approach whereby initial estimates of the source covariance matrices obtained via any source separation algorithm are post-processed by some smoothing technique so as to reduce artifacts. Several such techniques have been introduced in the context of speech denoising or beamforming [7,1,6,4,16] and employed for the post-filtering of linear source estimates in the context of determined audio source separation [9]. However, they have not yet been systematically studied in the context of under-determined source separation involving directional interference instead of a somewhat diffuse background. Also, most of these techniques are specifically designed for single-channel input. In the following, we propose multichannel extensions of three single-channel smoothing techniques [7,1,16] and compare them with two existing multichannel techniques [6,4] on a set of speech and music mixtures.

The structure of the paper is as follows. We explain how to initially estimate the source covariance matrices and present five multichannel smoothing techniques in Section 2. We assess the performance of each technique for various source separation algorithms in Section 3 and conclude in Section 4.

2 Source covariance estimation and smoothing

Let us denote by $\mathbf{x}(n,f)$ and $\mathbf{c}_j(n,f)$ the $I \times 1$ vectors of STFT coefficients of the mixture and the spatial image of source j respectively. We presume that estimates of the source spatial images or their parameters have been obtained via any

source separation algorithm and apply the following three-step post-processing. Firstly, assuming that $\mathbf{c}_j(n,f)$ follows a zero-mean Gaussian distribution with *local covariance matrix* $\mathbf{R}_{\mathbf{c}_j}(n,f)$ [14], we derive initial estimates $\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)$ of these covariance matrices. Secondly, we replace the classical *multichannel Wiener filter* [4,14]

$$\widehat{\mathbf{W}}_j(n,f) = \widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)\widehat{\mathbf{R}}_{\mathbf{x}}^{-1}(n,f), \tag{2}$$

where $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} \widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)$ is the estimated mixture covariance matrix, by a smooth filter $\widetilde{\mathbf{W}}_j(n,f)$. Finally, the source spatial images are recovered by

$$\widetilde{\mathbf{c}}_j(n,f) = \widetilde{\mathbf{W}}_j(n,f)\mathbf{x}(n,f). \tag{3}$$

In the following, we discuss the first two steps in more detail.

2.1 Initial source covariance estimation

Source separation algorithms can be broadly divided into two categories: linear *vs.* variance model-based algorithms [14]. Linear model-based algorithms such as binary masking [11] or ℓ_1-norm minimization [17] directly operate on the mixture STFT coefficients and provide estimates $\widehat{\mathbf{c}}_j(n,f)$ of the source spatial images. The source covariance matrices can then be naturally initialized as $\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f) = \widehat{\mathbf{c}}_j(n,f)\widehat{\mathbf{c}}_j^H(n,f)$ where ^{H} denotes Hermitian transposition. By contrast, variance-model based algorithms [14] represent the mixture by some parametric distribution and operate on the parameters of this distribution. Initial estimates of the sources covariances may then be derived from the estimated parameters. In the particular case when a Gaussian distribution is chosen [8,13], the source covariances are readily estimated as the output of the algorithm.

In both cases, we add a small regularization term $\epsilon\,\mathbf{I}$ to the initial covariance matrices, where $\mathbf{I}$ is the $I \times I$ identity matrix. This term ensures that the matrix inversions in (2), (5) and (8) can always be computed even when a single source is active. The regularization factor is set to $\epsilon = 10^{-6} \times \mathrm{tr}\,\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)$.

2.2 Spatial smoothing

A few multichannel smoothing techniques have been proposed in the beamforming literature in order to widen the spatial response of the Wiener filter, so as to reduce artifacts supposedly located close to the target source direction. While these techniques were originally formulated for a single source in the presence of background noise, their application to multiple sources is straightforward. One technique proposed in [4, eq. 55] amounts to interpolating the Wiener filter as

$$\widetilde{\mathbf{W}}_j^{\mathrm{SFS}}(n,f) = (1-\mu)\widehat{\mathbf{W}}_j(n,f) + \mu\,\mathbf{I}. \tag{4}$$

This is equivalent to time-domain interpolation of the estimated source spatial image signals with the mixture signal as suggested in [11]. Another technique stemming from a weighted likelihood model results in a distinct interpolation [6]

$$\widetilde{\mathbf{W}}_j^{\mathrm{SCS}}(n,f) = \widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)[(1-\mu)\widehat{\mathbf{R}}_{\mathbf{x}}(n,f) + \mu\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)]^{-1}. \tag{5}$$

We refer to the techniques in (4) and (5) as *spatial filter smoothing* (SFS) and *spatial covariance smoothing* (SCS) respectively. In both cases, the smoothness of the resulting filter increases with μ, so that it is equal to the conventional Wiener filter for $\mu = 0$ and to the identity filter for $\mu = 1$.

2.3 Temporal smoothing

Many techniques based on temporal smoothing of the source variances have also been proposed for single-channel speech denoising [16]. However, their extension to multichannel source separation is not straightforward. Two approaches may be taken: either split the source covariance matrices into a spectral power $v_j(n,f) = \mathrm{tr}\,\mathbf{R}_{\mathbf{c}_j}(n,f)$ and a spatial covariance matrix $\mathbf{R}_j(n,f) = v_j^{-1}(n,f)\mathbf{R}_{\mathbf{c}_j}(n,f)$ [13], process the spectral power alone via a single-channel technique and recombine it with the spatial covariance matrix, or design new smoothing equations that process spectral power and spatial covariance at the same time. Our preliminary experiments showed that the latter approach always performed better. Hence, we only present the new proposed smoothing equations below.

The most popular technique consists of smoothing the initial Signal-to-Noise Ratio (SNR) in a causal [7] or noncausal [5] fashion, with the latter resulting in better onset preservation. Numerous variants of this so-called decision-directed technique have been proposed [9]. By replacing variances by covariance matrices and ratios by matrix inversion, we extend it to source separation as

$$\begin{aligned}
\widehat{\mathbf{G}}_j(n,f) &= \widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)[\widehat{\mathbf{R}}_{\mathbf{x}}(n,f) - \widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)]^{-1} \\
\widetilde{\mathbf{G}}_j(n,f) &= \frac{1}{L+1}\sum_{l=-L/2}^{L/2} \widehat{\mathbf{G}}_j(n+l,f) \\
\widetilde{\mathbf{W}}_j^{\mathrm{TRS}}(n,f) &= \mathbf{I} - [\widetilde{\mathbf{G}}_j(n,f) + \mathbf{I}]^{-1}
\end{aligned} \tag{6}$$

where $\mathbf{G}_j(n,f)$ is a multichannel generalization of the SNR and we assume a noncausal rectangular smoothing window of length $L+1$. Note that this technique does not apply to binary masking, since $\widehat{\mathbf{G}}_j(n,f)$ is infinite in that case.

A simpler technique consists of smoothing the conventional single-channel Wiener filter [1], which easily extends to a multichannel setting as

$$\widetilde{\mathbf{W}}_j^{\mathrm{TFS}}(n,f) = \frac{1}{L+1}\sum_{l=-L/2}^{L/2} \widehat{\mathbf{W}}_j(n+l,f). \tag{7}$$

Finally, one may also compute the Wiener filter from smoothed source variances [16]. By smoothing the source covariances instead, we obtain

$$\begin{aligned}
\widetilde{\mathbf{R}}_{\mathbf{c}_j}(n,f) &= \frac{1}{L+1}\sum_{l=-L/2}^{L/2} \widehat{\mathbf{R}}_{\mathbf{c}_j}(n+l,f) \\
\widetilde{\mathbf{W}}_j^{\mathrm{TCS}}(n,f) &= \widetilde{\mathbf{R}}_{\mathbf{c}_j}(n,f)\widetilde{\mathbf{R}}_{\mathbf{x}}^{-1}(n,f)
\end{aligned} \tag{8}$$

with $\widetilde{\mathbf{R}}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} \widetilde{\mathbf{R}}_{\mathbf{c}_j}(n,f)$. We here consider sliding smoothing windows instead of disjoint windows as in [16].

We call the techniques in (6), (7) and (8) *temporal SNR smoothing* (TRS), *temporal filter smoothing* (TFS) and *temporal covariance smoothing* (TCS) respectively. Smoothness increases with L and the conventional Wiener filter is obtained for $L = 0$. Note that, contrary to spatial smoothing, the filter does not tend to identity when $L \to \infty$ but to a stationary Wiener filter instead.

3 Experimental evaluation

We applied the five above smoothing techniques for the post-processing of three separation algorithms, namely binary masking [11], ℓ_1-norm minimization with two active sources per time-frequency bin [17] and local Gaussian modeling [13], over four instantaneous stereo ($I = 2$) mixtures of $J = 3$ sources. These mixtures were taken from the 2008 Signal Separation Evaluation Campaign (SiSEC) [12] and cover both male and female speech, percussive and non-percussive music. *The mixing matrices were known.* Performance was evaluated using the overall Signal-to-Distortion Ratio (SDR) as well as the Signal-to-Interference Ratio (SIR) and the Signal-to-Artifacts Ratio (SAR) in [15], averaged over all sources and all mixtures. The choice of instantaneous mixing was dictated by the limited accuracy of these criteria in a convolutive setting. Indeed, while they are accurate for instantaneous mixtures, they do not yet provide sufficiently precise distinction of interference and artifacts on convolutive mixtures for this study.

The tradeoff between SAR and SIR as a function of μ and L is shown in Figure 1. Temporal covariance smoothing provides the best tradeoff independently of the considered separation algorithm. The resulting SIR decreases in similar proportion to the increase of the SAR and a small increase of the SIR is even observed for small μ or L. Spatial filter smoothing also improves the SAR but results in a much steeper decrease of the SIR. All other methods fail to reduce artifacts in a predictable manner and result either in non-monotonous increase or decrease of the SAR. This indicates that many state-of-the-art smoothing techniques initially designed for diffuse and/or stationary noise can fail in the presence of directional nonstationary sources. In particular, temporal SNR smoothing appears extremely sensitive to the initial estimation of the variances, while spatial covariance smoothing results in conventional Wiener filtering for all $0 \leq \mu < 1$ both for binary masking and ℓ_1-norm minimization[1] and in small deviation from conventional Wiener filtering for local Gaussian modeling.

These conclusions are further supported by the SDR curves in Figure 2, which decrease quickly for all techniques except for temporal covariance smoothing due to its good tradeoff between interference and artifacts and for spatial covariance smoothing as explained above. A SDR increase is even observed with temporal covariance smoothing, which is equal to 3 dB for binary masking and less for the two other algorithms.

[1] It can be shown that SCS amounts to conventional Wiener filtering for all $0 \leq \mu < 1$ when $\epsilon \to 0$ as soon as at most two sources are active in each time-frequency bin.

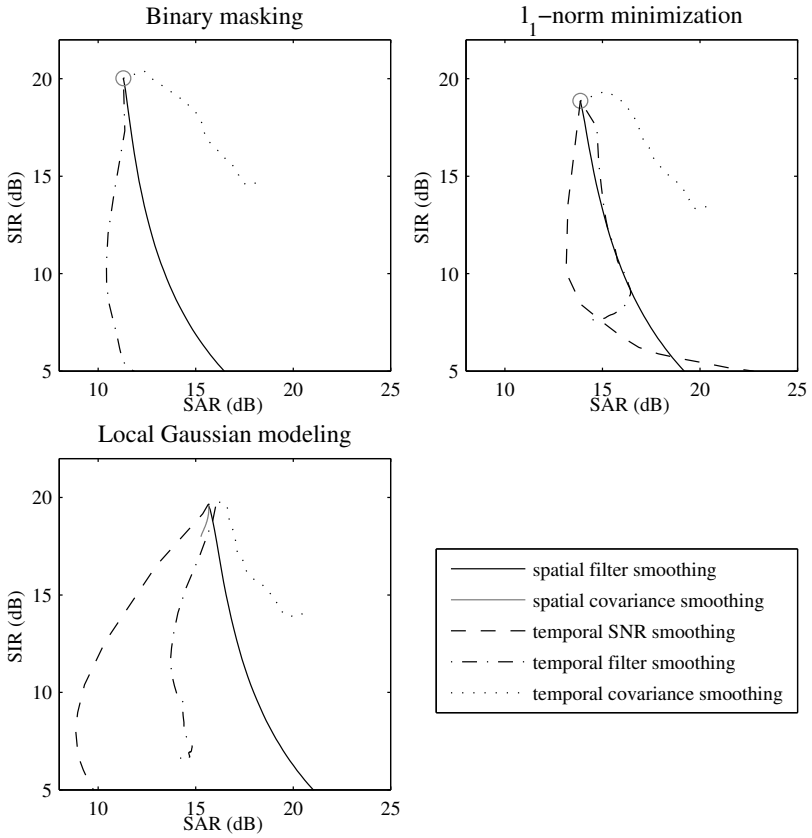

Fig. 1. Average tradeoff between SAR and SIR achieved by each separation algorithm and each smoothing technique.

4 Conclusion and perspectives

We reformulated state-of-the-art Wiener filter smoothing techniques in the context of under-determined audio source separation. Experimental results indicate that many techniques thought for spatially diffuse and/or stationary noise fail with directional nonstationary sources. Temporal covariance smoothing provides the best tradeoff between SAR and SIR and also potentially increases the overall SDR. Future work will concentrate on assessing robustness to mixing matrix estimation errors and adaptively estimating the optimal size L of the smoothing window in each time-frequency bin for that technique.

References

1. Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., et al.: Qualcomm-ICSI-OGI features for ASR. In: Proc. 7th Int. Conf. on Spoken Language Process-

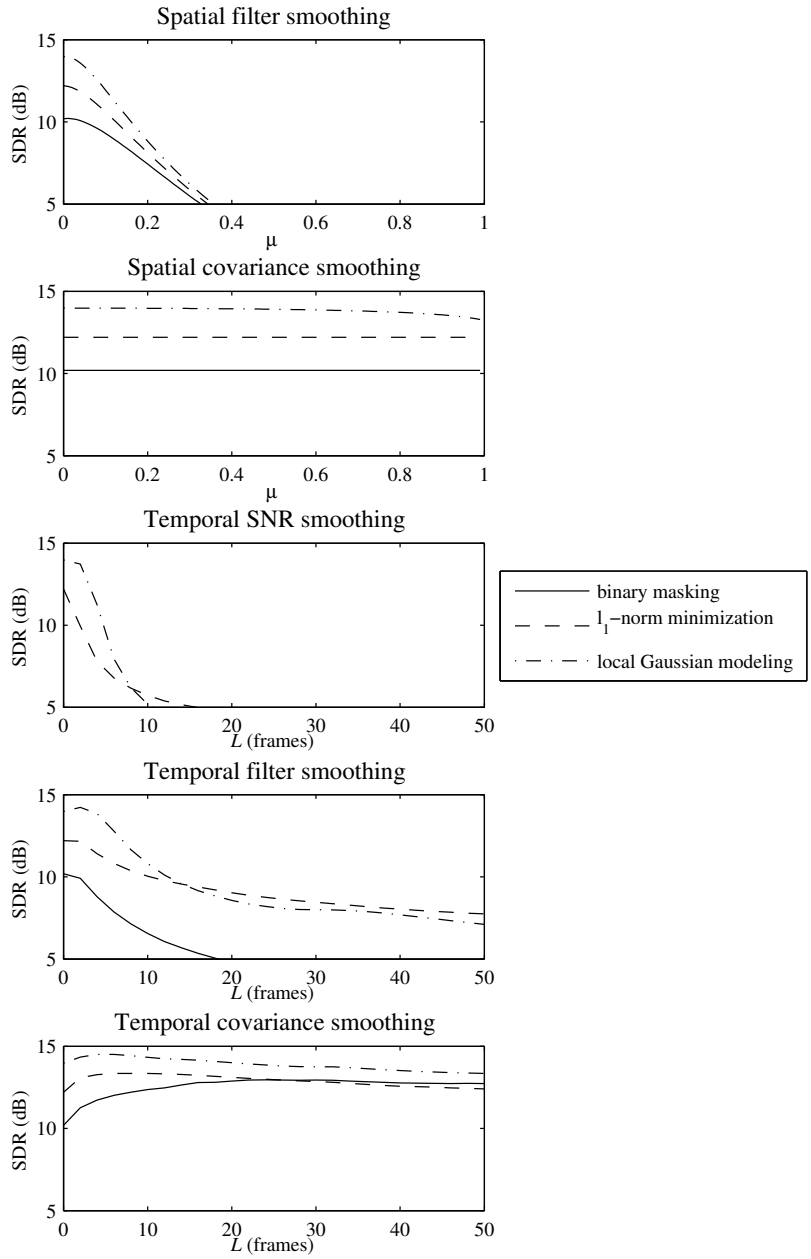

Fig. 2. Average SDR achieved by each separation algorithm and each smoothing technique as a function of the smoothing parameter.

ing. pp. 21–24 (2002)
2. Araki, S., Makino, S., Sawada, H., Mukai, R.: Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. In: Proc. 2005 IEEE Int. Conf. on Acoustics, Speech and Signal Processing. pp. III–81–III–84 (2005)
3. Araki, S., Sawada, H., Mukai, R., Makino, S.: Blind sparse source separation with spatially smoothed time-frequency masking. In: Proc. 2006 Int. Workshop on Acoustic Echo and Noise Control (2006)
4. Chen, J., Benesty, J., Huang, Y., Doclo, S.: New insights into the noise reduction Wiener filter. IEEE Transactions on Audio, Speech and Language Processing 14(4), 1218–1234 (2006)
5. Cohen, I.: Speech enhancement using a noncausal *a priori* SNR estimator. IEEE Signal Processing Letters 11(9), 725–728 (2004)
6. Doclo, S., Moonen, M.: On the output SNR of the speech-distortion weighted multichannel Wiener filter. IEEE Signal Processing Letters 12(12), 809–811 (2005)
7. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing 32(6), 1109–1121 (1984)
8. Févotte, C., Cardoso, J.F.: Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In: Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. pp. 78–81 (2005)
9. Hoffmann, E., Kolossa, D., Orglmeister, R.: Time frequency masking strategy for blind source separation of acoustic signals based on optimally-modified log-spectral amplitude estimator. In: Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation. pp. 581–588 (2009)
10. Nesbit, A., Jafari, M.G., Vincent, E., Plumbley, M.D.: Audio source separation using sparse representations. In: Machine Audition : Priciples, Algorithms and Systems. IGI Global (2010)
11. Rickard, S.T.: The DUET blind source separation algorithm. In: Blind Speech Separation, pp. 217–237. Springer (2007)
12. Vincent, E., Araki, S., Bofill, P.: The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation. In: Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA). pp. 734–741 (2009)
13. Vincent, E., Arberet, S., Gribonval, R.: Underdetermined instantaneous audio source separation via local Gaussian modeling. In: Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation. pp. 775–782 (2009)
14. Vincent, E., Jafari, M.G., Abdallah, S.A., Plumbley, M.D., Davies, M.E.: Probabilistic modeling paradigms for audio source separation. In: Machine Audition : Priciples, Algorithms and Systems. IGI Global (2010)
15. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P.: First stereo audio source separation evaluation campaign: Data, algorithms and results. In: Proc. 7th Int. Conf. on Independent Component Analysis and Signal Separation. pp. 552–559 (2007)
16. Yu, G., Mallat, S., Bacry, E.: Audio denoising by time-frequency block thresholding. IEEE Transactions on Signal Processing 56(5), 1830–1839 (2008)
17. Zibulevsky, M., Pearlmutter, B.A., Bofill, P., Kisilev, P.: Blind source separation by sparse decomposition in a signal dictionary. In: Independent Component Analysis: Principles and Practice, pp. 181–208. Cambridge Press (2001)