Consistent Wiener filtering for audio source separation

Jonathan Le Roux, Emmanuel Vincent

Consistent Wiener Filtering for Audio Source Separation

Jonathan Le Roux, *Member, IEEE,* and Emmanuel Vincent, *Senior Member, IEEE*

***Abstract*—Wiener filtering is one of the most ubiquitous tools in signal processing, in particular for signal denoising and source separation. In the context of audio, it is typically applied in the time-frequency domain by means of the short-time Fourier transform (STFT). Such processing does generally not take into account the relationship between STFT coefficients in different time-frequency bins due to the redundancy of the STFT, which we refer to as *consistency*. We propose to enforce this relationship in the design of the Wiener filter, either as a hard constraint or as a soft penalty. We derive two conjugate gradient algorithms for the computation of the filter coefficients and show improved audio source separation performance compared to the classical Wiener filter both in oracle and in blind conditions.**

***Index Terms*—Wiener filtering, Short-time Fourier transform, Spectrogram consistency, Source separation, Conjugate gradient**

EDICS Category: SAS-ICAB

I. Introduction

WIENER filtering is one of the most widely used tools in signal processing, in particular for signal denoising and source separation. In the context of audio, where signals are not stationary but short-term stationary, it is typically applied in the time-frequency domain via the short-time Fourier transform (STFT). The filter coefficients, which are nonnegative, are independently computed in each time-frequency bin from the variances of the signals to be separated without accounting for the *consistency* of the obtained set of STFT coefficients, i.e., the fact that they actually correspond to the STFT of a time-domain signal [1]. Together with inaccurate estimation of the signal variances, this independent processing is one of the reasons behind the presence of residual interference or musical noise in the separated signals [2].

One way to promote consistency of the separated STFT coefficients is to employ phase reconstruction methods, which estimate consistent phases given magnitude spectra [1], [3]. Multiple input spectrogram inversion (MISI) [4] and partitioned phase retrieval (PPR) [5] methods were specifically designed to handle multiple sources. Because they modify phase only, these methods implicitly assume that the input magnitude spectra are close to the true source spectra. This assumption is valid in the context of informed source separation [6] but not in the usual context of blind or semi-blind source separation of interest here, where both the estimated source variances and the magnitude spectra resulting from Wiener filtering may differ from the true source spectra. As a result, these methods provide limited improvement as we shall see in the following.

In earlier work [7], we proposed a complex-valued variant of the Wiener filter promoting consistency of the separated STFT coefficients via a soft penalty term. In contrast to the above methods, the magnitude and phase of the coefficients are jointly optimized and the resulting filter retains all properties of the classical Wiener filter, namely minimum mean square error (MMSE) optimality under Gaussian assumptions and exact reconstruction of the mixture from the separated sources. Although the problem can be formulated as the minimization of a quadratic function, exact inversion of the associated linear system is typically intractable. In [7], we proposed an iterative auxiliary function-based algorithm for mixtures of two sources which converged slowly for large values of the penalty weight parameter, so that a dynamical update scheme for this parameter had to be introduced. Another algorithm for the joint modification of magnitude and phase was recently proposed in [6] under the name of informed source separation using iterative reconstruction (ISSIR).

We present here two new optimization algorithms based on the conjugate gradient method [8] with well-chosen preconditioners, which enforce consistency either as a *hard* constraint or as a *soft* penalty for any number of sources. We show that, for a suitable choice of the penalty weight, the latter provides better separation performance than the former. It also outperforms the algorithms in [4]–[7] both in oracle conditions, when the variances of the sources are known, and in blind conditions, when they are estimated from the noisy input.

After presenting a general maximum-likelihood formulation of the problem in Section II, we describe the proposed algorithms in Sections III and IV, and evaluate their performance through audio source separation experiments in Section V.

II. Wiener filtering and consistency

Let us consider an observed single-channel audio signal $x(t)$ that is a mixture of J sound sources $s_j(t)$,

$$x(t) = \sum_{j=1}^{J} s_j(t), \tag{1}$$

where J is known and t denotes discrete time. This is equivalently written in the time-frequency domain as

$$X_{nf} = \sum_{j} S_{jnf}. \tag{2}$$

Assuming that the STFT coefficients S_{jnf} of the sources are independent zero-mean Gaussian random variables with positive variance v_{jnf} in each time frame n and frequency bin f, the STFT coefficients X_{nf} of the mixture also follow independent zero-mean Gaussian distributions with variance

$$v^x_{nf} = \sum_{j=1}^{J} v_{jnf}. \tag{3}$$

We are interested in the conditional distribution of the sources S_{jnf} given the mixture X_{nf} in each time-frequency bin (n, f). Due to the constraint (2), this conditional distribution lies on a subspace of dimension $J-1$ so we focus on a subset of free variables. Without loss of generality, we consider the first $J' = J-1$ sources as free variables given the mixture and denote them as $\boldsymbol{S}_{nf} = [S_{1nf}, \ldots, S_{J'nf}]^T$. Since $\boldsymbol{S}_{nf}$ and X_{nf} are jointly Gaussian with zero mean, their conditional distribution $p(\boldsymbol{S}_{nf}|X_{nf})$ is also Gaussian and its mean $\hat{\boldsymbol{\mu}}_{nf}$ and covariance $\boldsymbol{\Lambda}^{-1}_{nf}$ are given by (see for example [9])

$$\hat{\boldsymbol{\mu}}_{nf} = \boldsymbol{\Sigma}_{\boldsymbol{S}X} \boldsymbol{\Sigma}^{-1}_{XX} X_{nf}, \tag{4}$$

$$\boldsymbol{\Lambda}^{-1}_{nf} = \boldsymbol{\Sigma}_{\boldsymbol{SS}} - \boldsymbol{\Sigma}_{\boldsymbol{S}X} \boldsymbol{\Sigma}^{-1}_{XX} \boldsymbol{\Sigma}_{X\boldsymbol{S}}. \tag{5}$$

J. Le Roux was with NTT Communication Science Laboratories. He is now with Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA (e-mail: jonathan.le-roux@normalesup.org).

E. Vincent is with INRIA, Centre de Rennes - Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France (e-mail: emmanuel.vincent@inria.fr).

Here, $\mathbf{\Sigma}_{X\boldsymbol{S}} = \mathbf{\Sigma}_{\boldsymbol{S}X}^T = [v_{1nf}, \ldots, v_{J'nf}]$, $\mathbf{\Sigma}_{XX} = v_{nf}^x$, and $\mathbf{\Sigma}_{\boldsymbol{SS}}$ is the diagonal matrix with entries $v_{1nf}, \ldots, v_{J'nf}$. The conditional mean simplifies in particular to

$$\hat{\boldsymbol{\mu}}_{nf} = \left[\frac{v_{1nf}}{v_{nf}^x}, \ldots, \frac{v_{J'nf}}{v_{nf}^x}\right]^T X_{nf}, \tag{6}$$

which is none other than the classical Wiener filter output, i.e., the MMSE estimate of the sources. Using Woodbury's identity on Eq. (5), we obtain a simple expression for the precision $\mathbf{\Lambda}_{nf}$:

$$\mathbf{\Lambda}_{nf} = \begin{bmatrix} \frac{1}{v_{1nf}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{v_{J'nf}} \end{bmatrix} + \frac{1}{v_{Jnf}} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}. \tag{7}$$

We shall consider in the following the negative log-likelihood of the conditional distribution $-\log P(\boldsymbol{S}|X)$, which is equal up to a constant to the quadratic loss function

$$\psi(\boldsymbol{S}) = \sum_{n,f} (\boldsymbol{S}_{nf} - \hat{\boldsymbol{\mu}}_{nf})^H \mathbf{\Lambda}_{nf} (\boldsymbol{S}_{nf} - \hat{\boldsymbol{\mu}}_{nf}), \tag{8}$$

where $\boldsymbol{S}$ and X respectively denote the sets of all variables $\boldsymbol{S}_{nf}$ and X_{nf}. Without further constraint on $\boldsymbol{S}$, the maximum-likelihood solution is obviously the classical Wiener filter output $\boldsymbol{S} = \hat{\boldsymbol{\mu}}$.

However, we need to keep in mind that the STFT, when classically computed using overlapping windows, is a redundant representation which results in a certain relationship between the coefficients in different time-frequency bins. Denoting by N the number of time frames and F the number of frequency bins, let us consider the set $\mathcal{S}$ of arrays $\boldsymbol{S}$ of $\mathbb{C}^{J' \times N \times F}$ such that $S_{jn,F-f} = \bar{S}_{jnf}$ for all j, n and f. The STFT of J' real-valued time-domain signals is an element of $\mathcal{S}$, which we call consistent, but not all elements of $\mathcal{S}$ can be obtained as such [1], [3]. If the inverse STFT (iSTFT) is defined in such a way that time-domain signals can be exactly reconstructed from their STFT through iSTFT without additional time-varying normalization, a necessary and sufficient condition for an element $\boldsymbol{S}$ of $\mathcal{S}$ to be consistent is that it is equal to the STFT of its iSTFT [1] or, in other words, that it belongs to the null space $\mathrm{Ker}\,\mathcal{F}$ of the $\mathbb{R}$-linear operator $\mathcal{F}$ from $\mathcal{S}$ to itself defined by

$$\mathcal{F} = \mathrm{Id} - \mathrm{STFT} \circ \mathrm{iSTFT} \tag{9}$$

where Id denotes the identity operator and the STFT and iSTFT operators are separately applied to each signal. The classical Wiener filter output $\boldsymbol{S} = \hat{\boldsymbol{\mu}}$ does generally not satisfy this constraint, so that the STFT of the resynthesized signals $\mathrm{iSTFT}(\hat{\boldsymbol{\mu}})$ differs from $\hat{\boldsymbol{\mu}}$ and no longer maximizes the conditional log-likelihood.

III. Consistency as a hard constraint

A first natural approach to enforcing consistency in the Wiener filter is to minimize $\psi(\boldsymbol{S})$ under the constraint that $\mathcal{F}(\boldsymbol{S}) = 0$. This is equivalent to finding time-domain source signals $\boldsymbol{s}(t) = [s_1(t), \ldots, s_{J'}(t)]^T$ minimizing $\psi(\mathrm{STFT}(\boldsymbol{s}))$.

Setting the gradient of $\psi(\mathrm{STFT}(\boldsymbol{s}))$ with respect to $\boldsymbol{s}$ to 0, we see that its global minimum verifies

$$\mathrm{STFT}^* \circ \mathbf{\Lambda} \circ \mathrm{STFT}(\boldsymbol{s}) = \mathrm{STFT}^* \circ \mathbf{\Lambda}(\hat{\boldsymbol{\mu}}) \tag{10}$$

where STFT^* is the adjoint of the STFT and the operator $\mathbf{\Lambda}$ denotes multiplication by the matrix $\mathbf{\Lambda}_{nf}$ in each time-frequency bin (n, f), i.e., $\mathbf{\Lambda}(\mathbf{S})_{nf} = \mathbf{\Lambda}_{nf}\mathbf{S}_{nf}$. Under the common assumption that the synthesis window of the iSTFT is equal to the analysis window of the STFT up to a multiplicative constant, it can be shown that $\mathrm{STFT}^* = M\,\mathrm{iSTFT}$ where M is the frame length [10]. The global minimum is therefore equivalently given by

$$\mathrm{iSTFT} \circ \mathbf{\Lambda} \circ \mathrm{STFT}(\boldsymbol{s}) = \mathrm{iSTFT} \circ \mathbf{\Lambda}(\hat{\boldsymbol{\mu}}). \tag{11}$$

Algorithm 1 Conjugate gradient with hard constraint

Inputs: $\mathbf{\Lambda}$, $\hat{\boldsymbol{\mu}}$, $\epsilon > 0$
$\boldsymbol{s}_0 \leftarrow \mathrm{iSTFT}(\hat{\boldsymbol{\mu}})$
$\boldsymbol{r}_0 \leftarrow \mathrm{iSTFT} \circ \mathbf{\Lambda}(\hat{\boldsymbol{\mu}}) - \mathrm{iSTFT} \circ \mathbf{\Lambda} \circ \mathrm{STFT}(\boldsymbol{s}_0)$
$\boldsymbol{p}_0 \leftarrow \mathrm{iSTFT} \circ \mathbf{\Lambda}^{-1} \circ \mathrm{STFT}(\boldsymbol{r}_0)$
$\delta_{\text{new}} \leftarrow \langle \boldsymbol{r}_0, \boldsymbol{p}_0 \rangle$
$k \leftarrow 0$
repeat
  $\boldsymbol{q}_k \leftarrow \mathrm{iSTFT} \circ \mathbf{\Lambda} \circ \mathrm{STFT}(\boldsymbol{p}_k)$
  $\alpha_k \leftarrow \frac{\delta_{\text{new}}}{\langle \boldsymbol{p}_k, \boldsymbol{q}_k \rangle}$
  $\boldsymbol{s}_{k+1} \leftarrow \boldsymbol{s}_k + \alpha_k \boldsymbol{p}_k$
  $\boldsymbol{r}_{k+1} \leftarrow \boldsymbol{r}_k - \alpha_k \boldsymbol{q}_k$
  $\boldsymbol{z}_{k+1} \leftarrow \mathrm{iSTFT} \circ \mathbf{\Lambda}^{-1} \circ \mathrm{STFT}(\boldsymbol{r}_{k+1})$
  $\delta_{\text{old}} \leftarrow \delta_{\text{new}}$
  $\delta_{\text{new}} \leftarrow \langle \boldsymbol{r}_{k+1}, \boldsymbol{z}_{k+1} \rangle$
  $\beta_k \leftarrow \frac{\delta_{\text{new}}}{\delta_{\text{old}}}$
  $\boldsymbol{p}_{k+1} \leftarrow \boldsymbol{z}_{k+1} + \beta_k \boldsymbol{p}_k$
  $k \leftarrow k + 1$
until $\alpha_{k-1}^2 \|\boldsymbol{p}_{k-1}\|^2 < \epsilon \|\boldsymbol{s}_k\|^2$
return $\boldsymbol{s}_k$

Although the invertibility of the Hermitian operator $\mathrm{iSTFT} \circ \mathbf{\Lambda} \circ \mathrm{STFT}$ seems difficult to prove, we believe that it always holds in practice. For it to be singular, there would need to exist signals whose STFTs, once multiplied in each bin by the real-valued matrix $\mathbf{\Lambda}_{nf}$, would become perfectly inconsistent despite the lack of phase modification. We believe this situation to be unlikely to exist, and indeed always observed finite condition numbers in our experiments (see Fig. 1).

A solution to the linear system (11) can be iteratively estimated through the preconditioned conjugate gradient method [8]. Because the operator $\mathbf{\Lambda}$ is typically badly conditioned, $\mathrm{iSTFT} \circ \mathbf{\Lambda} \circ \mathrm{STFT}$ is itself badly conditioned and the choice of a suitable preconditioner $\boldsymbol{M}$ is crucial for fast convergence. We propose to choose the preconditioner as

$$\boldsymbol{M}^{-1} = \mathrm{iSTFT} \circ \mathbf{\Lambda}^{-1} \circ \mathrm{STFT} \tag{12}$$

where $\mathbf{\Lambda}^{-1}$ denotes multiplication by the matrix $\mathbf{\Lambda}_{nf}^{-1}$ in each time-frequency bin (n, f), such that $\boldsymbol{M}^{-1}$ approximates the inverse of the operator to be inverted.

The resulting algorithm is described in Algorithm 1, with $\langle ., . \rangle$ denoting the dot product. The computational complexity of each iteration is dominated by four STFT or iSTFT operations.

IV. Consistency as a soft penalty

As an alternative approach, we consider relaxing the hard consistency constraint $\mathcal{F}(\boldsymbol{S}) = 0$, which may be inadequate when the estimated source variances are unreliable, by introducing the L^2 norm of $\mathcal{F}(\boldsymbol{S})$ as a soft penalty term in (8) with some weight γ. We thus obtain a new objective function:

$$\psi_\gamma(\boldsymbol{S}) = \psi(\boldsymbol{S}) + \gamma \sum_{n,f} \left\| \mathcal{F}(\boldsymbol{S})_{nf} \right\|^2. \tag{13}$$

If the weight of the penalty is chosen sufficiently large, the estimated spectrograms should finally both be consistent and minimize ψ among all consistent spectrograms.

Setting the gradient of $\psi_\gamma(\boldsymbol{S})$ with respect to $\boldsymbol{S}$ to 0, we see that its global minimum verifies

$$(\mathbf{\Lambda} + \gamma \mathcal{F}^* \circ \mathcal{F})(\boldsymbol{S}) = \mathbf{\Lambda}(\hat{\boldsymbol{\mu}}). \tag{14}$$

It can easily be shown that $\mathcal{F}$ is a projector, i.e., $\mathcal{F} \circ \mathcal{F} = \mathcal{F}$, and that, under the same assumption on synthesis and analysis windows

Algorithm 2 Conjugate gradient with soft penalty

Inputs: $\boldsymbol{\Lambda}$, $\hat{\boldsymbol{\mu}}$, $\gamma > 0$, $\epsilon > 0$
$\boldsymbol{S}_0 \leftarrow \hat{\boldsymbol{\mu}}$
$\boldsymbol{R}_0 \leftarrow -\gamma\mathcal{F}(\boldsymbol{S}_0)$
$\boldsymbol{P}_0 \leftarrow (\boldsymbol{\Lambda} + \gamma\frac{FN-T}{FN}\mathrm{Id})^{-1}(\boldsymbol{R}_0)$
$\delta_{\text{new}} \leftarrow \langle \boldsymbol{R}_0, \boldsymbol{P}_0 \rangle$
$k \leftarrow 0$
repeat
  $\boldsymbol{Q}_k \leftarrow \boldsymbol{\Lambda}(\boldsymbol{P}_k) + \gamma\mathcal{F}(\boldsymbol{P}_k)$
  $\alpha_k \leftarrow \frac{\delta_{\text{new}}}{\langle \boldsymbol{P}_k, \boldsymbol{Q}_k \rangle}$
  $\boldsymbol{S}_{k+1} \leftarrow \boldsymbol{S}_k + \alpha_k \boldsymbol{P}_k$
  $\boldsymbol{R}_{k+1} \leftarrow \boldsymbol{R}_k - \alpha_k \boldsymbol{Q}_k$
  $\boldsymbol{Z}_{k+1} \leftarrow (\boldsymbol{\Lambda} + \gamma\frac{FN-T}{FN}\mathrm{Id})^{-1}(\boldsymbol{r}_{k+1})$
  $\delta_{\text{old}} \leftarrow \delta_{\text{new}}$
  $\delta_{\text{new}} \leftarrow \langle \boldsymbol{R}_{k+1}, \boldsymbol{Z}_{k+1} \rangle$
  $\beta_k \leftarrow \frac{\delta_{\text{new}}}{\delta_{\text{old}}}$
  $\boldsymbol{P}_{k+1} \leftarrow \boldsymbol{Z}_{k+1} + \beta_k \boldsymbol{P}_k$
  $k \leftarrow k + 1$
until $\alpha_{k-1}^2 \|\boldsymbol{P}_{k-1}\|^2 < \epsilon \|\boldsymbol{S}_k\|^2$
return $\boldsymbol{S}_k$

as in Section III, it is Hermitian, i.e., $\mathcal{F}^* = \mathcal{F}$. As a consequence, $\mathcal{F}^* \circ \mathcal{F} = \mathcal{F}$, and (14) becomes

$$(\boldsymbol{\Lambda} + \gamma\mathcal{F})(\boldsymbol{S}) = \boldsymbol{\Lambda}(\hat{\boldsymbol{\mu}}). \tag{15}$$

We use the preconditioned conjugate gradient method to invert (15). As the operator $\boldsymbol{\Lambda} + \gamma\mathcal{F}$ is ill-conditioned, the choice of a good preconditioner is crucial again. Denoting by T the signal length, the projector $\mathcal{F}$ can be diagonalized on two eigenspaces: its null space $\mathrm{Ker}\,\mathcal{F}$ for the eigenvalue 0 with multiplicity $J'\,T$ and its image space $\mathrm{Im}\,\mathcal{F}$ for the eigenvalue 1 with multiplicity $J'\,(FN-T)$. On average, the eigenvalues of $\mathcal{F}$ are thus equal to $\frac{FN-T}{FN}$. A suitable choice for the preconditioner $\boldsymbol{M}$ is then

$$\boldsymbol{M}^{-1} = \left(\boldsymbol{\Lambda} + \gamma\frac{FN-T}{FN}\mathrm{Id}\right)^{-1}, \tag{16}$$

where $\boldsymbol{M}^{-1}$ amounts to binwise matrix multiplication.

The resulting algorithm is described in Algorithm 2. The computational complexity of each iteration is dominated by two STFT or iSTFT operations, which is comparable to the complexity of the algorithm in [7] or half that of Algorithm 1.

V. Experimental evaluation

A. Setup

We now evaluate the separation performance of the proposed algorithms on mixtures of single-channel speech and real-world non-stationary noise signals taken from the development set of the 2010 Signal Separation Evaluation Campaign (SiSEC) for the task "Source separation in the presence of real-world background noise" [2]. The noise signals were recorded in a subway, on a square and in a cafeteria. For each of nine pairs of speech and noise signals in the SiSEC data, we created mixtures at three input Signal to Noise Ratios (SNR): -10, 0 and $+10$ decibels (dB). In order to avoid dependency of the algorithm behavior on the scale of $\boldsymbol{\Lambda}$, each mixture signal was normalized so that its root mean square amplitude was equal to 0.063. All signals were sampled at 16 kHz and had a duration of 10 s. The STFT and iSTFT were implemented as in [11] and computed using half-overlapping sine windows of length $M = 1024$.

The soft penalty-based algorithm was implemented with a fixed weight γ, for $\gamma \in \{10^0, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$. The algorithm using a hard constraint will be formally denoted by $\gamma = +\infty$. The stopping criterion was chosen as $\epsilon = 10^{-6}$, so that the separation performance is similar to that obtained by exact inversion of (11) and (15), as verified by preliminary experiments with much smaller ϵ. Note that ϵ is the sole parameter of both algorithms, while the algorithm in [7] required additional parameters to be set for the dynamic update scheme of the penalty weight. For comparison, we also evaluated the classical Wiener filter output $\hat{\boldsymbol{\mu}}$, which corresponds to $\gamma = 0$, our previous method [7], PPR [5], and the result of applying MISI [4] and single-resolution ISSIR [6] to the magnitude of $\hat{\boldsymbol{\mu}}$, using the same stopping criterion as above and default values for the other parameters as given in [4]–[7]. Note that stopping some of the methods prior to convergence may lead to a shorter computation time with similar performance [5], but we preferred to use the same convergence criterion for all methods to ease comparison.

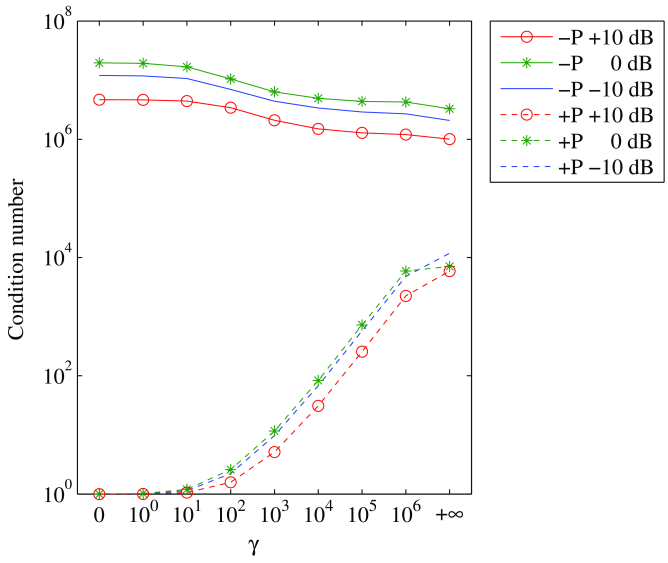

Fig. 1. Condition numbers of the linear systems (11) and (15) in oracle conditions with $(+P)$ and without $(-P)$ preconditioning, for the classical Wiener filter ($\gamma = 0$) and for the proposed constraint-based ($\gamma = +\infty$) and penalty-based algorithms, averaged over all mixtures for each input SNR.

We assessed all algorithms in two different conditions: an *oracle* condition yielding an upper bound on performance, in which the source variances are set to the true power STFTs of the speech and noise signals, and a *blind* condition, in which only the stationary noise spectrum is assumed to be known and fixed to the time average of the power STFT of the noise signal and the speech variance is estimated from the mixture by means of spectral subtraction [12]. For each algorithm, we evaluated the separation performance via the overall Signal-to-Distortion ratio (SDR) and further analyzed it by means of the Signal-to-Interference Ratio (SIR) and Signal-to-Artifacts Ratio (SAR) metrics [2]. We also report the computation time using Matlab on a 2.93 GHz quad core CPU.

B. Results

In Fig. 1, we show the average condition numbers of the linear systems (11) and (15) with and without preconditioning in the oracle case (the blind case leads to similar numbers). This quantity is equal to the ratio of the largest and smallest eigenvalues of the operators to be inverted, which can be computed via the well-known power iteration algorithm. For all values of γ, the proposed preconditioners reduce the condition numbers by several orders of magnitude and hence lead to a considerable convergence speedup [8].

In Fig. 2, we show the SDR achieved by the proposed constraint-based and penalty-based algorithms as a function of γ. We see that

TABLE I

Comparison of the SDR, SIR, SAR (dB) and computation time (s), averaged over all mixtures for each input SNR, for the classical Wiener filter, MISI [4], ISSIR [6], PPR [5], our previous method [7], and the proposed penalty-based algorithm with γ chosen by cross-validation for each input SNR, in oracle and blind conditions.

	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 [4]	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 [5]	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 [6]	11.7	22.6	12.3	2.5	17.9	25.7	18.8	1.3	24.8	30.4	26.4	0.7
	Le Roux et al. [7]	12.8	24.6	13.6	5.2	19.0	27.3	20.0	5.0	25.7	31.4	27.3	4.0
	Penalty (Proposed)	12.6	25.2	13.5	2.6	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]	-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 [5]	-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 [6]	-2.3	-0.8	1.0	17.8	6.6	10.0	9.3	7.8	15.8	18.7	19.1	2.5
	Le Roux et al. [7]	-1.2	0.6	0.9	13.7	7.5	12.1	9.3	11.9	16.6	20.9	18.8	9.5
	Penalty (Proposed)	0.5	3.3	1.2	10.6	8.6	12.3	10.9	3.4	17.1	20.6	20.0	2.1

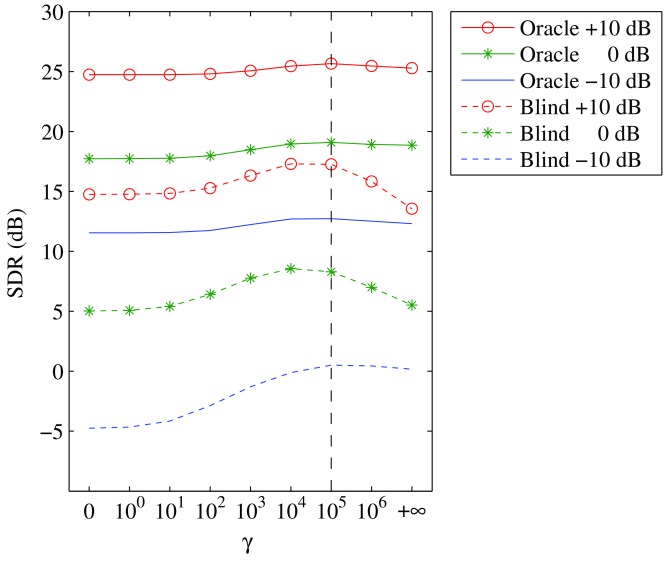

Fig. 2. Output SDR of the classical Wiener filter ($\gamma = 0$) and the proposed constraint-based ($\gamma = +\infty$) and penalty-based algorithms in oracle and blind conditions, averaged over all mixtures for each input SNR.

the soft penalty-based approach with $\gamma = 10^5$ is generally the best overall, especially in blind conditions when the noise and speech variances are inaccurately estimated.

Numerical results for the classical Wiener filter, MISI, ISSIR, PPR, our previous method and the proposed penalty-based algorithm are given in Table I. The weight γ in the proposed algorithm is set for each file and each input SNR by cross-validation, using the weight that led to highest average SDR on all other files (note that similar results are obtained with γ fixed to 10^5). These confirm that the proposed method leads to a significant improvement over the classical Wiener filter and all previous algorithms in terms of SDR for both oracle and blind conditions, and to similar or better performance compared to our previous algorithm [7]. In the oracle case, the SIR and SAR also improve, while in the blind case the increase of the SDR can be further analyzed as a strong improvement of the SIR, offset by a small deterioration of the SAR. Informal listening tests support this conclusion in terms of the perceptual salience of interference and artifacts. The perceptual quality improvement is particularly noticeable for the SNR of -10 dB in the oracle case and for the SNR of 0 dB in the blind case. MISI and PPR did not lead to improvement in blind conditions, while ISSIR led to a smaller improvement than the proposed method. The computation time is in almost all cases shorter than the signal duration. The proposed method is faster than [7] for better or similar performance, and runs comparably to other methods but with better performance.

We also tested MISI and ISSIR on the magnitude spectrograms obtained by spectral subtraction in blind conditions, leading to worse performance by MISI and slightly better for ISSIR while still lower than our method (results not shown in the table).

VI. Conclusion

We showed how to account for the redundancy of the STFT in Wiener filtering and presented efficient optimization algorithms based on the conjugate gradient method with suitable preconditioning. The proposed algorithms lead to significant improvements of source separation performance under both oracle and blind conditions. Future work will deal with the extension of these algorithms to the multichannel case.

References

[1] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. SAPA*, Sep. 2008, pp. 23–28.

[2] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[3] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. ASSP*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[4] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Sig. Proc. letters*, vol. 17, no. 5, pp. 421–424, May 2010.

[5] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Proc. ICASSP*, Mar. 2012, pp. 101–104.

[6] ——, "Informed source separation using iterative reconstruction," *submitted to IEEE Trans. ASLP*, 2012, arXiv:1202.2075v1.

[7] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama, "Consistent Wiener filtering: Generalized time-frequency masking respecting spectrogram consistency," in *Proc. LVA/ICA*, Sep. 2010, pp. 89–96.

[8] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," CMU, Tech. Rep., 1994.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006, pp. 85–87.

[10] B. Yang, "A study of inverse short-time Fourier transform," in *Proc. ICASSP*, Apr. 2008, pp. 3541–3544.

[11] E. Vincent, R. Gribonval, and M. D. Plumbley, "BSS Oracle Toolbox Version 2.1," http://bass-db.gforge.inria.fr/bss oracle/.

[12] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol. 27, pp. 113–120, Apr. 1979.