MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Consistent Anisotropic Wiener Filtering for Audio Source

Magron, P.; Le Roux, J.; Virtanen, T.

TR2017-151 October 2017

Abstract

For audio source separation applications, it is common to apply a Wiener-like filtering to a time-frequency (TF) representation of the data, such as the short-time Fourier transform (STFT). This approach, which boils down to assigning the phase of the original mixture to each component, is limited when sources overlap in the TF domain. In this paper, we propose a more sophisticated version of this technique for improved phase recovery. First, we model the sources by anisotropic Gaussian variables: this model accounts for the non-uniformity of the phase, and then permits us to incorporate some prior information about the phase that originates from a sinusoidal model. Then, we exploit the STFT consistency, which is the relationship between STFT coefficients that is due to its redundancy. We derive a conjugate gradient algorithm for estimating the corresponding filter, called consistent anisotropic Wiener. Experiments conducted on music pieces show that accounting for those two phase properties outperforms each approach taken separately.

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)

CONSISTENT ANISOTROPIC WIENER FILTERING FOR AUDIO SOURCE SEPARATION

Paul Magron,[1] *Jonathan Le Roux,*[2] *Tuomas Virtanen,*[1] *

[1] Signal Processing Laboratory, Tampere University of Technology (TUT), Finland
paul.magron@tut.fi, tuomas.virtanen@tut.fi
[2] Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA
leroux@merl.com

ABSTRACT

For audio source separation applications, it is common to apply a Wiener-like filtering to a time-frequency (TF) representation of the data, such as the short-time Fourier transform (STFT). This approach, which boils down to assigning the phase of the original mixture to each component, is limited when sources overlap in the TF domain. In this paper, we propose a more sophisticated version of this technique for improved phase recovery. First, we model the sources by anisotropic Gaussian variables: this model accounts for the non-uniformity of the phase, and then permits us to incorporate some prior information about the phase that originates from a sinusoidal model. Then, we exploit the STFT consistency, which is the relationship between STFT coefficients that is due to its redundancy. We derive a conjugate gradient algorithm for estimating the corresponding filter, called consistent anisotropic Wiener. Experiments conducted on music pieces show that accounting for those two phase properties outperforms each approach taken separately.

***Index Terms*—** Wiener filtering, phase recovery, sinusoidal modeling, STFT consistency, audio source separation.

1. INTRODUCTION

Audio source separation consists in extracting underlying components called *sources* that add up to form an observable audio signal called *mixture*. Many audio source separation techniques act in the time-frequency (TF) domain, which reveals the particular structure of these signals. Most techniques, such as nonnegative matrix factorization (NMF) [1], are applied to nonnegative-valued representations (e.g. short-time Fourier transform (STFT) spectrograms), because the structure of sound is more prominent in that domain. They have shown promising for audio source separation [2, 3].

However, when it comes to resynthesizing time-domain signals, obtaining the phase of the corresponding complex-valued STFT is necessary [4, 5]. In the single-channel source separation framework, a common practice consists in applying a Wiener-like filtering [3], which boils down to assigning the phase of the mixture to each extracted component. Such a filter, which is optimal in a minimum mean square error (MMSE) sense, originates from the underlying assumption that the phase is uniformly-distributed [6]. It has however been pointed out [7] that Wiener filtering fails to provide good results when sources overlap in the TF domain, thus highlighting the need for novel phase recovery techniques.

In this way, consistency-based approaches can be used for phase recovery [8]. That is, a complex-valued matrix is iteratively computed in order to maximize its consistency, i.e., to bring it as close as possible to the STFT of a time signal, as this property is not guaranteed by phase retrieval methods in general. Some recent works [9, 10, 11] attempted to combine Wiener filtering and consistency-based techniques in a unified framework for audio source separation. Consistent Wiener filtering [11] has been shown to be the most promising candidate for this task.

Alternatively, phase recovery can be performed by using phase models based on signal analysis. For instance, the widely used model of mixtures of sinusoids [12] leads to explicit constraints for phase reconstruction that are based on the relationships between adjacent TF bins [13]. Such an approach has been exploited in the phase vocoder algorithm [14] for time-stretching, speech enhancement [15, 16], audio restoration [13] and source separation [17]. In [18], we introduced an anisotropic Gaussian (AG) model in which the phase is no longer uniform, which permits us to incorporate some prior information about the phase that arises from a sinusoidal model. We derived an MMSE estimator which generalizes Wiener filtering to AG variables.

In this paper, we propose to combine those two approaches by exploiting both a consistency constraint and some phase information based on a signal model. We propose to address this issue by extending the consistent Wiener filtering to the AG case. Our approach then consists in minimizing an objective cost function which penalizes the reconstruction error in the AG model, to which is added a regularization term which promotes consistency. This function is minimized by means of the preconditioned conjugate gradient algorithm. Experiments conducted on realistic music signals for a vocals/accompaniment separation task show that exploiting those two phase constraints within a unified framework outperforms both approaches taken separately.

This paper is organized as follows. Section 2 presents the generalized anisotropic Wiener filtering and details the estimation of the sources under a consistency constraint. Section 3 experimentally validates the potential of this method for an audio source separation task. Finally, Section 4 draws some concluding remarks.

2. CONSISTENT ANISOTROPIC WIENER FILTERING

2.1. Anisotropic Gaussian model

Let $X \in \mathbb{C}^{F \times T}$ be the STFT of a single-channel audio signal. X is the linear and instantaneous mixture of J sources S_j, such that for all TF bin ft:

$$X_{ft} = \sum_j S_{jft}. \qquad (1)$$

[*] This work was partly supported by the Audiobrain project funded by the Academy of Finland.

Since all TF bins are treated independently, we remove the indices f and t in what follows for more clarity. We assume that each source S_j follows a complex normal distribution: $S_j \sim \mathcal{N}(m_j, \gamma_j, c_j)$, where m_j (resp. γ_j and c_j) is the mean (resp. the variance and the relation term) of S_j. The covariance matrix is defined as:

$$\Gamma_j = \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix}, \tag{2}$$

where $\bar{z}$ denotes the complex conjugate of z. Many previous studies [3, 11, 19] model the sources as circular-symmetric (or *isotropic*) variables [3] (i.e., such that $m_j = c_j = 0$), which boils down to assuming that the phase of each source is uniformly-distributed.

In this paper, we adopt a different standpoint, originally developed in [18]: we model the source signals by mixtures of sinusoids, which leads to explicit relationships between the phases of adjacent TF bins [13] and therefore to some prior phase estimate ϕ_j. We then consider that the phases should be distributed around the values ϕ_j with a concentration parameter $\kappa \in]0, +\infty[$. Thus, we propose to structure the moments of the distribution as follows[1]:

$$m_j = \lambda\sqrt{v_j}e^{i\phi_j},\ \gamma_j = (1-\lambda^2)v_j \text{ and } c_j = \rho v_j e^{i2\phi_j}, \tag{3}$$

where $\lambda = \dfrac{I_1(\kappa)}{I_0(\kappa)}$, $\rho = \dfrac{I_2(\kappa)}{I_0(\kappa)} - \lambda^2$, I_n is the modified Bessel function of the first kind of order n and v_j is an estimate of the source power $|S_j|^2$ obtained beforehand (e.g. after a preliminary NMF [1]). The relation terms c_j are non-zero in general, which conveys the property of *anisotropy* of the corresponding Gaussian distribution: this is why we refer to it as the anisotropic Gaussian (AG) model.

The additive property of the Gaussian distribution family then implies that $X \sim \mathcal{N}(m_X, \gamma_X, c_X) = \mathcal{N}(\sum_j m_j, \sum_j \gamma_j, \sum_j c_j)$, and $\Gamma_X = \sum_j \Gamma_j$.

2.2. MMSE estimation without constraint

We seek to obtain an estimator of the sources for performing the separation task. We consider the posterior distribution of the sources given the mixture. Due to the constraint (1), this conditional distribution lies on a subspace of dimension $J' = J-1$ so we focus on a subset of free variables. Without loss of generality, we consider the first J' sources as free variables given the mixture and denote them as $\mathbf{S} = [S_1, ..., S_{J'}]^T$ in each TF bin ft, where $.^T$ denotes the transpose. It can be shown [20] that $\mathbf{S}|X$ follows a multivariate complex normal distribution with mean vector $\boldsymbol{\mu} = [\mu_1, ..., \mu_{J'}]^T$ such that:

$$\underline{\mu}_j = \underline{m}_j + \Gamma_j \Gamma_X^{-1}(\underline{X} - \underline{m}_X), \tag{4}$$

where $\underline{x} = \begin{pmatrix} x & \bar{x} \end{pmatrix}^T$. The posterior covariance matrix is:

$$\Xi = \begin{pmatrix} \Gamma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Gamma_{J'} \end{pmatrix} - \begin{pmatrix} \Gamma_1 \\ \vdots \\ \Gamma_{J'} \end{pmatrix} \Gamma_X^{-1} \begin{pmatrix} \Gamma_1 \\ \vdots \\ \Gamma_{J'} \end{pmatrix}^T. \tag{5}$$

In particular, the posterior covariance matrix of each source is $\Gamma'_j = \Gamma_j - \Gamma_j \Gamma_X^{-1} \Gamma_j$. Using the Woodbury identity, we obtain the precision matrix Λ defined as the inverse of the covariance matrix:

$$\Lambda = \Xi^{-1} = \begin{pmatrix} \Gamma_1^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Gamma_{J'}^{-1} \end{pmatrix} + \begin{pmatrix} \Gamma_J^{-1} & \dots & \Gamma_J^{-1} \\ \vdots & \ddots & \vdots \\ \Gamma_J^{-1} & \dots & \Gamma_J^{-1} \end{pmatrix}. \tag{6}$$

Therefore, the negative log-likelihood of the posterior distribution $-\log p(\mathbf{S}|X)$ is equal, up to an additive constant and to a positive scale factor, to the following quadratic loss function:

$$\Psi(S) = \sum_{ft} (\underline{\mathbf{S}_{ft}} - \underline{\boldsymbol{\mu}_{ft}})^H \Lambda_{ft} (\underline{\mathbf{S}_{ft}} - \underline{\boldsymbol{\mu}_{ft}}), \tag{7}$$

where $\underline{\mathbf{S}} = [S_1, \bar{S_1}, ..., S_{J'}, \bar{S_{J'}}]^T$ (the notation $\underline{\boldsymbol{\mu}}$ is similar) and $.^H$ denotes the conjugate transpose. Setting the gradient of Ψ in (7) w.r.t. $\mathbf{S}_{ft}$ to 0 leads to the MMSE solution: $S_{jft} = \mu_{jft}, \forall j, f, t$.

2.3. Consistency constraint

When the STFT is computed using overlapping analysis windows (which is usual in practice), it is a redundant TF representation which implies that certain relationships must hold between its TF coefficients. This results in the fact that not all matrices in $\mathbb{C}^{F\times T}$ are the STFT of a time-domain signal. We will then say that a matrix is *consistent* [21] if it is equal to the STFT of its inverse STFT, or, in other words, if it belongs to Ker($\mathcal{F}$), where:

$$\forall S \in \mathbb{C}^{F\times T}, \mathcal{F}(S) = S - \text{STFT} \circ \text{iSTFT}(S). \tag{8}$$

The Wiener filter output does not generally satisfy this constraint, so that STFT ∘ iSTFT(μ) no longer minimizes the loss function (7).

We focus on the case $J = 2$ (i.e., $J' = 1$). This corresponds to many source separation applications where only 2 sources interact, such as speech/noise or singing voice/musical accompaniment. Moreover, the general case can be reduced to this special case by considering in turn each source against all others. Since in this case $\mathbf{S}_{ft}$ reduces to S_{1ft}, we shall remove the index $j = 1$ for clarity. The cost function (7) then rewrites:

$$\Psi(S) = \sum_{f,t} (\underline{S}_{ft} - \underline{\mu}_{ft})^H \Lambda_{ft} (\underline{S}_{ft} - \underline{\mu}_{ft}), \tag{9}$$

where $\Lambda_{ft} = \Gamma_{1ft}^{-1} + \Gamma_{2ft}^{-1} = \Gamma'^{-1}_{ft}$. As in [11], we propose to promote consistency in the form of a soft penalty added to the cost (9), which results in the following new objective function:

$$\Psi_\delta(S) = \Psi(S) + 2\delta||\mathcal{F}(S)||^2, \tag{10}$$

where $||.||$ denotes the Frobenius norm for matrices. The greater δ, the more consistent the resulting source estimate will be.

We can find the complex spectrogram[2] S minimizing Ψ_δ by setting the gradient of $\Psi_\delta(S)$ to 0. The consistency term is identical to that in [11], but the gradient of $\Psi(S)$ is here slightly more involved. To make its derivation easier to understand, it helps to consider the whole complex spectrogram S as the equivalent vector $\vec{S}$ obtained by concatenating the real and imaginary parts of all the frames of S. The gradient of $\Psi(S)$ can be derived with respect to the elements of $\vec{S}$, leading to an $\mathbb{R}$-linear operator on $\vec{S} - \vec{\mu}$, which can be reformulated as an $\mathbb{R}$-linear operator on $S - \mu$. We eventually obtain the

[1] The mathematical derivation of the moments can be obtained in [18].

[2] For convenience, we call "complex spectrogram" any complex-valued matrix, even if it is not the STFT of an actual signal.

gradient of $\Psi(S)$ w.r.t. S_{ft} as:

$$\nabla_{S_{ft}} \Psi(S) = 4\Omega_{ft}(S_{ft} - \mu_{ft}), \tag{11}$$

where

$$\Omega_{ft}(y) = \frac{1}{|\Gamma'_{ft}|}(\gamma'_{ft} y - c'_{ft}\bar{y}), \forall y \in \mathbb{C}, \tag{12}$$

and $|\Gamma'_{ft}| = \gamma'^2_{ft} - |c'_{ft}|^2$. Altogether, setting the gradient of $\Psi_\delta(S)$ to 0 leads to

$$(\Omega + \delta\mathcal{F}^* \circ \mathcal{F})S = \Omega\mu, \tag{13}$$

where * denotes the Hermitian adjoint and Ω is defined as the $\mathbb{R}$-linear operator that consists in independently applying Ω_{ft} to each TF bin Y_{ft} of a complex spectrogram Y:

$$(\Omega Y)_{ft} = \Omega_{ft}(Y_{ft}), \quad \forall Y \in \mathbb{C}^{F\times T}.$$

Since $\mathcal{F}$ is a projector, then $\mathcal{F} \circ \mathcal{F} = \mathcal{F}$. Furthermore, if the analysis and synthesis windows are equal up to a scaling factor (which is generally the case in practice), then it can be shown [11] that $\mathcal{F}$ is Hermitian, i.e., $\mathcal{F}^* = \mathcal{F}$. Therefore, $\mathcal{F}^* \circ \mathcal{F} = \mathcal{F}$, and the global minimum verifies:

$$(\Omega + \delta\mathcal{F})S = \Omega\mu. \tag{14}$$

Drawing on [11], we propose to solve (14) with the preconditioned conjugate gradient method [22], since the operator $\Omega + \delta\mathcal{F}$ is ill-conditioned. The preconditioner M is derived similarly to [11], leading here at each TF bin to

$$(MY)_{ft} = \Omega_{ft}(Y_{ft}) + \delta\frac{FT - L}{FT} Y_{ft}, \tag{15}$$

where L is the time-signal length. Inverting M is slightly more involved than in [11], where it amounted to a simple scalar multiplication, because $\Omega_{ft}(Y_{ft})$ here involves both Y_{ft} and $\bar{Y}_{ft}$ as can be seen in Eq. (12). A short calculation leads to

$$(M^{-1}(Y))_{ft} = \frac{1}{\eta_{ft}}\Big\{\Big(\frac{\gamma'_{ft}}{|\Gamma'_{ft}|} + \delta\frac{FT-L}{FT}\Big)Y_{ft} + \frac{c'_{ft}}{|\Gamma'_{ft}|}\bar{Y}_{ft}\Big\}, \tag{16}$$

where $\eta_{ft} = \left(\frac{\gamma'_{ft}}{|\Gamma'_{ft}|} + \delta\frac{FT-L}{FT}\right)^2 - \frac{|c'_{ft}|^2}{|\Gamma'_{ft}|^2}$.

The full procedure is summarized in Algorithm 1, and a MATLAB implementation is available at [23].

Algorithm 1 Consistent anisotropic Wiener filtering. Note: matrix operations are element-wise.

```
Inputs:
Posterior expectation $\mu \in \mathbb{C}^{F\times T}$,
Anisotropy and consistency parameters $\kappa \geq 0$ and $\delta \geq 0$,
Prior power $v \in \mathbb{R}_+^{2\times F\times T}$ and phase $\phi \in [0, 2\pi[^{2\times F\times T}$,
Stopping criterion $\epsilon > 0$.
Posterior moments
$\lambda = I_1(\kappa)/I_0(\kappa)$, $\rho = I_2(\kappa)/I_0(\kappa) - \lambda^2$.
$\gamma_1 = (1-\lambda^2)v_1$, $\gamma_2 = (1-\lambda^2)v_2$, $\gamma_X = \gamma_1 + \gamma_2$.
$c_1 = \rho v_1 e^{i2\phi_1}$, $c_2 = \rho v_2 e^{i2\phi_2}$, $c_X = c_1 + c_2$.
$\gamma' = \gamma_1 - \left(\gamma_X(\gamma_1^2 + |c_1|^2) - 2\gamma_1\Re(c_1\bar{c}_X)\right)/(\gamma_X^2 - |c_X|^2)$,
$c' = c_1 - \left(2\gamma_X\gamma_1 c_1 - \gamma_1^2 c_X - c_1^2\bar{c}_X\right)/(\gamma_X^2 - |c_X|^2)$,
$|\Gamma'| = \gamma'^2 - |c'|^2$.
Preconditioned conjugate gradient
$\Omega$ as defined in Eq. (12) and $M^{-1}$ as defined in Eq. (16),
$S_0 = \mu$,
$R_0 = -\delta\mathcal{F}(S_0)$,
$P_0 = M^{-1}(R_0)$,
$\xi_{\text{new}} = \langle R_0, P_0\rangle$,
$k = 0$.
repeat
    $Q_k = \Omega(P_k) + \delta\mathcal{F}(P_k)$,
    $\alpha_k = \xi_{\text{new}}/\langle P_k, Q_k\rangle$,
    $S_{k+1} = S_k + \alpha_k P_k$,
    $R_{k+1} = R_k - \alpha_k Q_k$,
    $Z_{k+1} = M^{-1}(R_{k+1})$,
    $\xi_{\text{old}} = \xi_{\text{new}}$,
    $\xi_{\text{new}} = \langle R_{k+1}, Z_{k+1}\rangle$,
    $\beta_k = \xi_{\text{new}}/\xi_{\text{old}}$,
    $P_{k+1} = Z_{k+1} + \beta_k P_k$,
    $k = k + 1$.
until $\alpha_{k-1}^2||P_{k-1}||^2 < \epsilon||S_k||^2$
Output:
$S_k \in \mathbb{C}^{F\times T}$.
```

3. EXPERIMENTAL EVALUATION

3.1. Dataset and protocol

We propose to experimentally assess the potential of the proposed consistent anisotropic Wiener filtering procedure described in Algorithm 1. We consider 100 music songs from the Demixing Secrets Database (DSD100), a remastered version of the database used for the SiSEC 2015 campaign [24]. The database is split into two sets of 50 songs: a training set and a test set. Each song is made up of $J = 2$ sources: the vocal track and the musical accompaniment track (which may contain various instruments such as guitar, bass, drums, piano...). The signals are sampled at $F_s = 44100$ Hz and the STFT is computed with a 46 ms long Hann window and 75 % overlap.

Two scenarios are considered: an Oracle scenario, in which the powers v are assumed to be known (i.e., equal to the ground truth), and an Informed scenario, in which the spectrograms are estimated from the Oracle values by means of an NMF with Kullback-Leibler divergence [2], which uses 100 iterations of multiplicative update rules and a rank of factorization $K = 50$. Note that this is not a fully blind scenario (it actually corresponds to the encoding stage in an informed source separation framework [19]), but this will inform us about the performance of the methods when the spectrograms differ from the ground truth.

When Algorithm 1 is initialized with Wiener filtering ($\phi = \angle X$ and $\kappa = 0$), we will refer to it as consistent Wiener filtering (CW). When it is initialized with anisotropic Wiener filtering (AW [18]), we will refer to it as consistent anisotropic Wiener filtering (CAW). As in [11], the stopping criterion is chosen as $\epsilon = 10^{-6}$.

The source separation quality is measured with scale-invariant versions [25] of the signal to distortion, interference and artifact ratios (SDR, SIR and SAR) [26] expressed in dB.

A demonstration on an audio excerpt is available at [23].

3.2. Influence of the consistency weight

First, similarly as in [18], we study the impact of the anisotropy parameter κ on the separation quality on the training set: the best results in terms of SDR, SIR and SAR are obtained for $\kappa = 1$ (resp. $\kappa = 0.8$) in the Oracle (resp. Informed) scenario.

Scenario	Method	Accompaniment			Singing voice			Avg. number
		SDR	SIR	SAR	SDR	SIR	SAR	of iterations
Oracle	Wiener [3]	16.7	27.2	17.2	12.1	26.7	12.3	-
	CW [11]	18.4	29.5	18.8	13.9	30.3	14.0	30
	AW [18]	17.5	27.9	17.9	13.0	28.0	13.1	-
	CAW (proposed)	**18.9**	**30.0**	**19.4**	**14.5**	**31.2**	**14.7**	**26**
Informed	Wiener [3]	15.9	26.1	16.4	11.3	25.7	11.5	-
	CW [11]	16.6	27.0	17.1	12.1	27.5	12.2	17
	AW [18]	16.1	26.3	16.6	11.5	26.2	11.7	-
	CAW (proposed)	**16.8**	**27.1**	**17.3**	**12.2**	**27.8**	**12.4**	**16**

Table 1: Average source separation performance for various methods on the DSD100 test dataset.

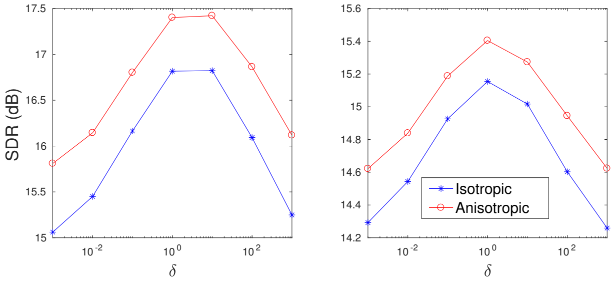

Figure 1: Influence of the consistency parameter δ on the source separation quality in Algorithm 1. The test is conducted in the Oracle (left) and Informed (right) scenarios.

We then investigate here the influence of the consistency parameter δ on the separation quality. The results in terms of SDR averaged over the 50 songs composing the training set are presented in Fig. 1 (similar trends are observed for the SIR and SAR).

We observe that promoting consistency leads to improving the separation quality over other approaches that do not account for this property (which correspond to $\delta = 0$), whether the magnitude values are known or estimated beforehand.

The optimal value of δ is dependent on the data, with a peak in the SDR here at 10 in the Oracle scenario and 1 in the Informed scenario. This corresponds to a trade-off between excessively promoting the consistency and only accounting for the MMSE estimates.

3.3. Separation results

We now consider the 50 songs that form the test set, and set δ to its learned optimal value. The results averaged over the dataset are presented in Table. 1.

In the Oracle scenario, the proposed method outperforms all the other approaches. While the AW technique leads to improving the separation quality over the Wiener estimates, it performs slightly worse than the CW filtering. The proposed CAW method overcomes this limit, since it combines the potential of both AW and CW approaches, and improves the criteria by approximately 0.5 dB over the CW technique. In the Informed scenario, the improvement is less significant (about 0.2 dB), which suggests that even if the proposed phase retrieval method can improve the separation quality over the other techniques, its full potential is reached when the power estimates are close to the ground truth.

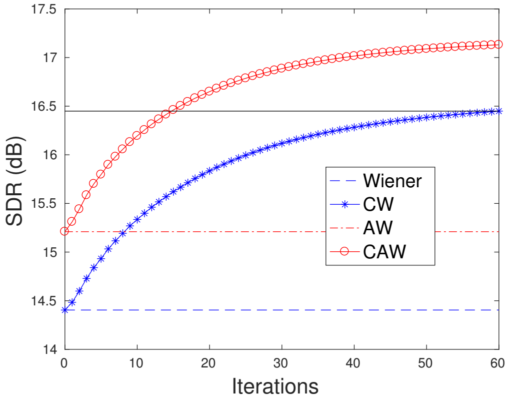

Figure 2: Separation quality (SDR in dB) over iterations.

Finally, the last column of Table 1 also indicates that CAW converges in less iterations on average than CW. We then show in Fig. 2 the evolution over iterations of the SDR, averaged over the test set in the Oracle scenario. For each excerpt, we run CW and CAW without using the stopping criterion, for 60 iterations (by 60 iterations, the algorithms had converged in all our experiments). We observe that the initial AW filtering approximately leads to the same results as 8 iterations of CW. Furthermore, 60 iterations of CW lead to a result (black solid line on the plot) that is similar to what is obtained with only 14 iterations of CAW. This shows that this anisotropic model, which accounts for a signal-based phase property, leads to a faster procedure than a phase-unaware source model. Overall, whether we look at the separation quality or the computational cost, CAW shows some improvement over the state-of-the-art CW.

4. CONCLUSION

The consistent anisotropic Wiener filtering procedure introduced in this paper is a promising approach for recovering the phase of the components in a source separation framework, since it combines a phase property that originates from signal modeling, and a consistency constraint which accounts for the redundancy of the STFT.

Future work will focus on extending this procedure to the case of more than 2 sources and to multichannel mixtures. In addition, such a technique can be implemented in an online fashion through a frame-by-frame processing, similarly as in some real-time implementations of the Griffin and Lim algorithm [27].

5. REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[3] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, March 2009.

[4] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, March 2015.

[5] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1 – 29, July 2016, phase-Aware Signal Processing in Speech Communication.

[6] R. M. Parry and I. Essa, "Incorporating Phase Information for Source Separation via Spectrogram Factorization," in *Proc. IEEE ICASSP*, vol. 2, Honolulu, Hawaii, USA, April 2007.

[7] P. Magron, R. Badeau, and B. David, "Phase recovery in NMF for audio source separation: an insightful benchmark," in *Proc. IEEE ICASSP*, Brisbane, Australia, April 2015.

[8] D. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, April 1984.

[9] D. Gunawan and D. Sen, "Iterative Phase Estimation for the Synthesis of Separated Sources From Single-Channel Mixtures," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421–424, May 2010.

[10] N. Sturmel and L. Daudet, "Informed Source Separation Using Iterative Reconstruction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 178–185, January 2013.

[11] J. Le Roux and E. Vincent, "Consistent Wiener Filtering for Audio Source Separation," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 217–220, Mar. 2013.

[12] R. J. McAuley and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, August 1986.

[13] P. Magron, R. Badeau, and B. David, "Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration," in *Proc. EUSIPCO*, Nice, France, August 2015.

[14] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, May 1999.

[15] P. Mowlaee and J. Kulmer, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1521–1532, September 2015.

[16] T. Gerkmann, "MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014.

[17] P. Magron, R. Badeau, and B. David, "Complex NMF under phase constraints based on signal modeling: application to audio source separation," in *Proc. IEEE ICASSP*, Shanghai, China, March 2016.

[18] ——, "Phase-dependent anisotropic Gaussian model for audio source separation," in *Proc. IEEE ICASSP*, New Orleans, LA, USA, March 2017.

[19] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[20] B. Picinbono, "Second-order complex random vectors and normal distributions," *IEEE Transactions on Signal Processing*, vol. 44, no. 10, pp. 2637–2640, Oct 1996.

[21] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. ISCA SAPA*, Brisbane, Australia, September 2008.

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

[23] http://www.cs.tut.fi/~magron/ressources/caw.tar.gz.

[24] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 387–395.

[25] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. ISCA Interspeech*, San Francisco, CA, USA, September 2016.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[27] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-Time Signal Estimation From Modified Short-Time Fourier Transform Magnitude Spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, 2007.