ITERATIVE PHASE RECONSTRUCTION OF WIENER FILTERED SIGNALS

Nicolas Sturmel, Laurent Daudet *

Institut Langevin
ESPCI, Univ. Paris Diderot, CNRS UMR 7587
10 rue Vauquelin 75005 Paris, France
{nicolas.sturmel,laurent.daudet}@espci.fr

ABSTRACT

This paper deals with phase estimation in the framework of underdetermined blind source separation, using an estimated spectrogram of the source and its associated Wiener filter. By thresholding the Wiener mask, two domains are defined on the spectrogram : a confidence domain where the phase is kept as the phase of the mixture, and its complement where the phase is updated with a projection similar to the widely-used Griffin and Lim technique. We show that with this simple technique, the choice of parameters results in a simple trade-off between distortion and interference. Experiments show that this technique brings significant improvements over the classical Wiener filter, while being much faster than other iterative methods.

Index Terms— Phase reconstruction, Spectrogram, STFT, Blind source separation, Wiener filter

1. INTRODUCTION

High-quality audio blind source separation, in the underdetermined case (more sources than sensors), is a very active and challenging topic. The separation itself often takes place in a domain where sources are sparse, usually in a time-frequency (TF) domain. The most popular choice of time-frequency transform is arguably the Short Time Fourier Transform (STFT), defined as:

$$S(m,n) = \sum_{k=0}^{L-1} w(k-mR)s(k)e^{-2j\pi nk/N},$$

where $s \in \mathbb{R}^L$ is the analyzed signal, w is the analysis window of support size L samples and window shift R samples, and N is the number of frequency bins. $S(m,n)$ is then a complex array of size (MxN) with Hermitian symmetry in frequency around $N/2$. For a well-chosen synthesis window, the STFT can be exactly inverted with standard overlap-add techniques. We denote $STFT^{-1}$ this inverse operator, so that $s = STFT^{-1}(STFT(s))$. Because STFTs are redundant representations, not every set $S \in \mathbb{C}^{M\times N}$, with the same Hermitian symmetry, represents a signal. For S to be a so-called consistent STFT, it must also verify the consistency equation:

$$S - \mathcal{G}(S) \quad = \quad 0 \tag{1}$$

Where the function $\mathcal{G}(S) = STFT[STFT^{-1}S]$ generates a consistent STFT from any set of complex values.

Consider the linear instantaneous mixture $x = \sum_i s_i$ of I sources. Source separation techniques try to estimate the individual sources s_i from the mixture. However, many of the more powerful methods, such as Non-negative Matrix Factorization (NMF [1]), only estimate the magnitude spectrogram W_i (the energy) of every source i in the time-frequency domain. This raises the following issues: (i) Only the source energy contribution to the mixture is known, and (ii) the phase of the source's TF distribution is unknown.

In order to tackle these problems, many solutions have been proposed, the best solution in the mean square sense being the widely used [2] Wiener Filtering. It masks the STFT of the mixture M with the real, positive coefficients $\alpha_i(m,n)$ such that :

$$\alpha_i(m,n) = \frac{W_i(m,n)}{\sum_k W_k(m,n)}$$

Then, the estimated STFT of the source $\hat{S}_i(m,n)$ is computed as $\hat{S}_i(m,n) = \alpha_i(m,n)X(m,n)$. This involves that the signal is reconstructed using the phase of the mixture. Wiener filter is mainly based on the separation of the sources in the spectral domain. When overlapping between sources increases, the reconstruction quality of the Wiener filter decreases. One solution to improve reconstruction quality, proposed by LeRoux et al. [3], is to constrain the estimation of $\hat{S}_i$ so that it satisfies the consistency constraint of equation (1) as well, at least approximately. However, this technique, called "consistent Wiener filtering" requires a careful balance between two terms, the Wiener filtering and the consistency, that must be dynamically updated throughout a sometimes large number of iterations. When the number of sources becomes large, and hence the overlap between sources increases, the tuning of these control parameters can become tricky.

In this paper, we resort to a somehow simpler method, where the parameters do not have to be adjusted on-line, and that therefore comes with a guaranteed small complexity. It is based on the popular phase reconstruction algorithm of Griffin and Lim (hereafter referred to as *G&L* [4]), here improved by a constraint based on the Wiener filter. The mechanic of such reconstruction will be discussed and compared to the "consistent Wiener filtering" both in terms of quality and complexity. It should be noted that for the sake of clarity and compactness, we present in this paper only the offline version of our algorithm (processing the signal as a whole), but the proposed modification can also be transposed to the newest real-time implementations [5, 6]. Note that related work [7] also uses a constrained version of G&L, however within a different framework.

The paper is organized as follows: we will first introduce the principle of iterative STFT phase reconstruction and the constrained Wiener filtering technique in the state of the art section 2. In section 3, we present our method and the experimental setup in section 4. We discuss the results in section 5 prior to the conclusion.

*This work was supported by the DReaM project (ANR-09-CORD-006) of the French National Research Agency CONTINT program. LD is on a joint position with Univ. Paris Diderot and Institut Universitaire de France

2. STATE OF THE ART

2.1. Iterative phase reconstruction

The consistency function given in equation (1) is used by the phase reconstruction algorithm proposed by G&L [4]. This technique aims at reconstructing the signal s while only knowing its spectrogram $W = |S|^2$. It computes the STFT $\tilde{S}$ which has the closest magnitude to the original spectrogram in the mean square sense. This method performs the update of equation (2) at each iteration k, applying both the consistency function $\mathcal{G}$ and a magnitude constraint.

$$\tilde{S}^{(k)} \quad = \quad \mathcal{G}(|S|e^{i\angle \tilde{S}^{(k-1)}}) \tag{2}$$

This simple method provides time-domain signals with good sound quality, although often with artifacts such as echo, smearing and modulations. However, its convergence is often painfully slow. Some recent studies [5, 6] significantly increased the speed of convergence, even allowing real-time online processing.

Unfortunately, this technique is not adapted, as such, to the source decomposition problem. Actually, since the estimated source STFT amplitude $|\hat{S}|$ is obtained by Wiener filtering, it does not represent the source S well enough to allow the method to convergence toward an optimal solution, as observed in [3]. Moreover, G&L minimization has a tendency to catch local minima such as translation or local inversion of the signal. See [8] for a complete state of the art.

2.2. Consistent Wiener filtering

The idea of combining Wiener filtering and consistency of the estimated spectrograms emerged in [3]. This method uses a consistency constraint while estimating a Wiener-like filter and is called *Consistent Wiener filtering*. This method uses an update of the form:

$$\hat{S}_i^{(p+1)} \quad \leftarrow \quad \frac{\frac{X}{\sum_{k\neq i} W_k} + \gamma \mathcal{G}(\hat{S}_i^{(p)})}{\frac{1}{W_i} + \frac{1}{\sum_{k\neq i} W_k} + \gamma} \tag{3}$$

Where $\hat{S}_i$ is the estimated source STFT and W_i is the spectrogram of the ith source. The parameter γ ($0 \leq \gamma$) weights the consistency constraint: $\gamma = 0$ is the standard Wiener estimate, while $\gamma \to \infty$ only enforces consistency. Setting this γ parameter is indeed crucial: we use the procedure given in [3] to dynamically update γ throughout the minimization process.

3. PARTITIONED PHASE RETRIEVAL

This paper proposes an alternative method for estimating the phase of the Wiener-filtered spectrogram $|\alpha_i X|^2$, while not modifying the amplitude. It is therefore closer to G&L's approach of the problem than consistent Wiener filtering. We will show that despite previous evidence, we can increase the source reconstruction quality with $G\&L$ by adding additional constraints.

Indeed, the G&L algorithm only constrains the magnitude of the STFT to be reconstructed. This can lead to local minima by the lack of phase information, but the Wiener filter does give additional information on the phase representation. Since the values of $\alpha_i(m,n) \in [0,1]$, the closer it is to 1, the closer is the mixture bin $X(m,n)$ to the original source bin $S_i(m,n)$, and so their respective phase. Therefore, one can decide of a confidence domain Ω_i:

$$\Omega_i \quad = \quad \{(m,n)|\alpha_i(m,n) > \tau\} \tag{4}$$

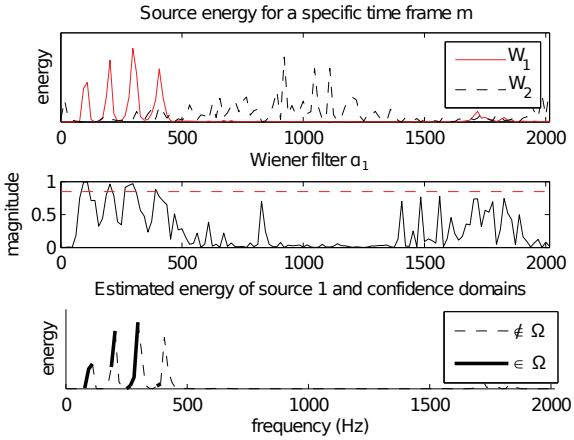

Fig. 1. Illustration of the confidence domain of the Wiener filter, on one frame m_0 of the spectrogram. Top : spectrogram amplitude of both sources, middle : coefficients $\alpha_1(m_0, n)$ of the Wiener mask for source 1, bottom : estimated spectrum of source 1, showing values where the mixture phase is kept (plain line), and values where phase is estimated through the G&L algorithm (dashed line).

The bins selected in Ω_i are mainly the bins of higher spectral energy of the source i. We consider that the phase value of each bin contained in Ω_i can be constrained as an accurate phase estimation of the target signal. This is especially true for harmonic sounds that are sparsely represented on the frequency axis, or percussive sounds that are sparsely represented on the time axis. Therefore the idea is not only to force the magnitude of the STFT, but also the phase of the Ω bins. This principle is presented on figure 1, where the energy of two sources is shown on top, the Wiener filter for source 1 is shown in the middle with the threshold τ, and the two different domains are shown at the bottom on the estimated Wiener spectrum.

As noted before, the estimated spectrogram W_i is not consistent: it does not correspond to the squared magnitude of the real source STFT S_i. Therefore, the "oracle" solution for the phase $\angle\hat{S}_i = \angle S_i$ does not lead to a perfect estimation of the source, but rather to the ideal phase reconstruction of our proposed method. Throughout the experiments we will refer to this solution as the "optimal" solution.

The method proposed here is to use Ω_i as a confidence domain in order to constrain the phase of the STFT. However, we also wish to enforce coherence of the STFT, as in [4, 6], and we use the function $\mathcal{G}$ previously defined. At the initialization stage, the STFT $\tilde{S}_i^{(0)}$ of the i-th source we are looking for is estimated as the Wiener estimate of the source: $\tilde{S}_i^{(0)} = \alpha X$

Then, for each iteration k we update $\tilde{S}_i$ with :

$$\tilde{S}_i^{(k+1)} \quad = \quad \begin{cases} |\alpha_i X|e^{j\angle\mathcal{G}(S_i^{(k)})} & \text{for } (m,n) \notin \Omega \\ \alpha_i X & \text{for } (m,n) \in \Omega \end{cases} \tag{5}$$

This process is repeated for a *small* number K of iterations: because this algorithm is initialized to the Wiener estimate of the source, the first step is already close to the optimum solution and therefore few iterations are necessary. The increase in quality can therefore always be relatively modest, but it is consistently positive.

This algorithm depends on two parameters: the threshold τ used on the Wiener mask to define the domain Ω_i, and the number K of iterations. Their influence can be shown on figure 2, on a mixture

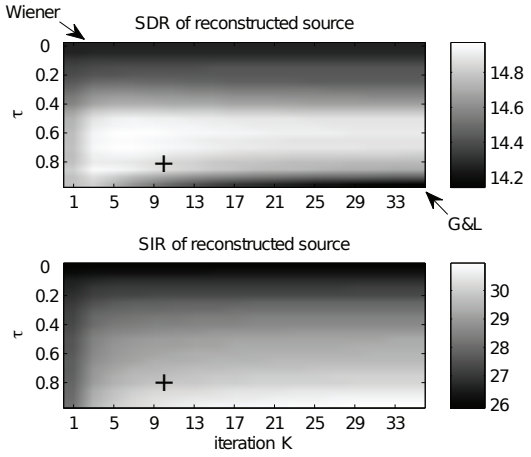

Fig. 2. Effect of the threshold τ and of the number K of iterations on SDR and SIR, on a speech mixture. The black cross indicates the choice of parameters used in the experimental evaluation. The G&L method and the standard Wiener filtering are also indicated, as special cases.

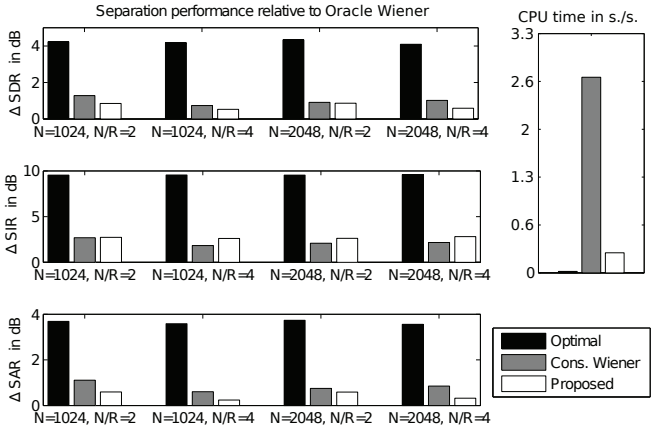

Fig. 3. Average separation performances for the three methods, on a mono mixture of 5 musical sources: upper bound with the optimal phase ("optimal"), Consistent Wiener, and the proposed method. All results are differential, in comparison to the baseline Oracle Wiener filter. Computation times are also given.

of two speech signals from the TIMIT database sampled at 16kHz for a window size of 1024 samples and 50% overlap. We analyze two objective quality criteria: the Signal to Distortion Ratio (SDR) and the Signal to Interference Ratio (SIR) from the bss eval toolbox [9]. As illustrated on figure 2, the higher τ, the higher the iterations, the better the SIR but not the SDR. Only optimizing the SDR leads to an optimum value of $\tau = 0.65$ and $K = 8$ iterations. This can be understood as follows: the SDR is the spectral distortion between the target source S and the estimated source $\tilde{S}$. In fact, the minimization we propose does not recover the real source: it estimates the most consistent phase pattern that suits to the STFT amplitude of the Wiener filter and some selected phase bins of the mixture. Because the target spectrogram $|\alpha_i X|^2$ is not consistent, and because the G&L algorithm naturally presents issues preventing convergence (stagnations [8]), improving the number of iterations does not necessarily improve the distortion, but it can still improve the interferences. Those stagnations can augment the absolute error between the reconstructed source and the original while only marginally lowering the perceptual quality.

When τ is too close to 1, there is not enough information to constrain the reconstruction (we get close to the classical G&L) and when τ gets close to 0, we get actually closer to the original Wiener solution. As a tradeoff between SIR and SDR, we arbitrarily choose $\tau = 0.8$ and $K = 10$ iterations. This allows very fast computation as we will see in the experiments of the next section. It should be emphasized that these experiments show that the results do not depend strongly on the exact values of the parameters ; these can be fixed once and for all depending on the desired SIR/SDR tradeoff. Note that the signals used for this parameter tuning were different from the ones used for the more extensive tests of the next section.

4. EXPERIMENTAL EVALUATION

Following the methodology in previous research, we tested this method in oracle conditions, with perfect knowledge of the energy of each sources composing the signal. We test music mixtures in the form of monophonic linear instantaneous mixes of 4 to 5 sources. We used three different extract of 10 to 15s length (a piece of Electro Jazz, Roxanne from the group Police and Call Me from the group Blondie) containing various instruments.

In order to assess our results we used the three objective criterions : SDR and SIR presented before, an the Source to Artifact Ratio (SAR). Because we use the Wiener filter as a reference, we will mainly display improvements of each methods, for this we use ΔSDR, ΔSIR and ΔSAR. The three values are defined as the mean improvement over the Wiener filter estimate for each source. Figure 3 shows the results of the experiment in the form of the mean improvement over the Oracle Wiener filter. We compare the proposed method to the solution where the phase is perfectly known ("Optimal"), and to the consistent Wiener filtering (Cons. Wiener) detailed in section 2.2 with all parameters as described in [3], including the number of iterations. In order to evaluate the effect of the redundancy and stationarity assumption of the TF representation, we test two window lengths and two overlap values (N/R). Separation enhancement are also compared to the respective computation time on figure 4 for the N=2048, N/R=2 case. Times are given in second of CPU time per second of signal. A demo page[1] allows audition of selected reconstructed samples.

5. DISCUSSION

The first striking result is the net gain on the SIR in comparison to both the Wiener filter solution and the consistent Wiener filter solution. This is a key point: we show here that it is possible to lower interferences from other sources by simply adjusting the phase profile of the estimated source STFT at a very small computational cost. Compared to classical Wiener filtering, the proposed method also performs better on the distortion and artifact components, but to a lesser extent. Optimal results show the upper limit of the Wiener-based phase reconstruction approach, due to the artifacts present in the magnitude STFT of each source.

No method really benefits from higher temporal overlap of the TF representation. When the overlap increases, the number of coefficients per time frame of the signal increases, but those coefficients

[1] http://nicolas.sturmel.com/ICASSP12

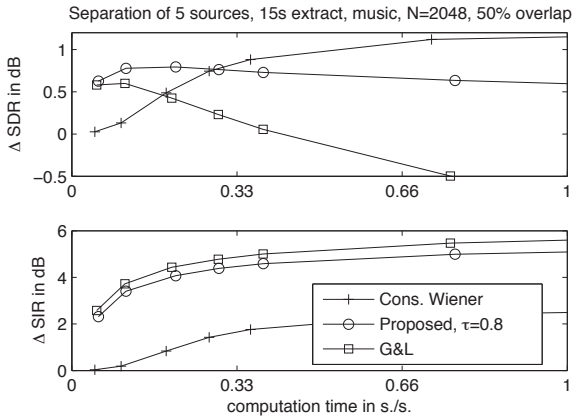

Fig. 4. Computation time in seconds of CPU per second of signal for a 15s music extract (Electro Jazz - 5 sources) for the Consistent Wiener, the classical G&L reconstruction, and our proposed method. Both ΔSIR and ΔSDR are given in reference to the separation performance of the classical Wiener filter estimate of each source.

are still badly estimated by the Wiener filter. In this case, more redundancy in the TF representation does not bring any new information useful for the reconstruction.

It has been shown in [8] that G&L was prone to catch local minima, especially because of the fundamental sign invariance of the solution (x or -x are equally valid solutions). This is especially true for "local" patches of the TF representation where no phase information is constrained. Therefore, G&L can improve the SIR and create artifacts at the same time: the reconstructed phase profile can be as far as the original signal (lowering the SAR) as it is far from the other sources (increasing SIR). Our method partially corrects this.

Auditory inspection of the separation as is possible on the demo page, shows that despite more audible artifacts than the consistent Wiener filtering, interferences are clearly lowered by our method. Lowering the interferences can sometimes be more important than lowering artifacts, especially when compensating those artifact with a remixing constraint as in [7].

At its best results, the proposed method is 7 to 10 times faster than the consistent Wiener filtering, because we perform much less iterations, with a lower complexity per iterate. In the proposed algorithms, both parameters are fixed so that we did not have to resort to ad-hoc adjustments needed on γ for the consistent Wiener. On figure 4 one can see that for a similar computation time, our method always outperforms the consistent Wiener filtering in terms of SIR, but also outperforms in terms of SDR for small computation times below 0.3 second of CPU per s. of signal. This makes our method best suited to real time implementations. In confirmation of [3] we can observe that for a higher number of iterations the simple G&L reconstruction generates a lot of artifacts ($\Delta SDR < 0$ with $\Delta SIR > 0$), more than it improves the interference. The key point of our method is then to provide similar improvements in interference rejection than G&L while generating much less artifacts, actually improving the Wiener estimation in every way.

6. CONCLUSION

A new method enhancing the source separation in the spectrogram domain has been presented, based on the Griffin and Lim phase reconstruction. This method adds an additional constraint on the bins using a thresholded Wiener filter: the phase is trusted at time-frequency bins where the Wiener mask is high, and estimated elsewhere. The corresponding algorithm can be easily parametrized according to the desired balance between SIR and SDR.

Results show that this approach consistently increases a plain application of the Wiener filter. The results are somewhat similar to those obtained by the Constrained Wiener filtering from LeRoux et al. [3]: the Constrained Wiener appears slightly better on SAR, and the proposed method is slightly better on SIR. However, the new approach has two important benefits. Firstly, it does not require a dynamical update of the parameters, which makes it suitable for a wider variety of test cases without *ad-hoc* tuning, especially in the case of a high number of sources leading to higher spectral overlap. Secondly, since the parameters are fixed, only a few iterations are necessary to get good results: the computational cost of this algorithm can be kept very low without sacrificing the output quality. Note that embedding the Wiener filter with the mixture to allow active listening is already done in [10]. Such method could benefit of such real-time ready separation enhancement.

Future work will focus on two key points: first, the adaptation of the algorithm to a real-time framework described in [5]: the low computational cost should easily allow the algorithm to run in real time. Then, the performance of this algorithm should be assessed for real (non-oracle, or only partially informed) source separation problems including perceptual evaluation.

7. REFERENCES

[1] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *WASPAA 2003*, 2003, pp. 177–180.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE TASSP*, vol. 32, pp. 1109–1121, dec 1984.

[3] 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 2010*, Sept. 2010, pp. 89–96.

[4] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE TASSP*, vol. 32(2), pp. 236–243, 1984.

[5] X. Zhu, G. Beauregard, and L. Wyse, "Real-time signal estimation from modified short-time fourier transform magnitude spectra," *IEEE TASLP*, vol. 15, no. 5, pp. 1645 –1653, 2007.

[6] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency," *proceedings of DAFx'10*, 2010.

[7] 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.

[8] N. Sturmel and L. Daudet, "Signal reconstruction from stft magnitude : a state of the art," in *proc. of DAFx'11, Paris*, 2011.

[9] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462 –1469, july 2006.

[10] H.-O. Oh, Y.-W. Jung, A. Favrot, and C. Faller, "Enhancing stereo audio with remix capability," in *In AES 129th Convention*, 2010.