



HAL
open science

Spectrogram Inversion for Audio Source Separation via Consistency, Mixing, and Magnitude Constraints

Paul Magron, Tuomas Virtanen

► **To cite this version:**

Paul Magron, Tuomas Virtanen. Spectrogram Inversion for Audio Source Separation via Consistency, Mixing, and Magnitude Constraints. European Signal Processing Conference (EUSIPCO), European Association for Signal Processing (EURASIP), Sep 2023, Helsinki, Finland. hal-04015991v1

HAL Id: hal-04015991

<https://hal.science/hal-04015991v1>

Submitted on 6 Mar 2023 (v1), last revised 17 Oct 2023 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Spectrogram Inversion for Audio Source Separation via Consistency, Mixing, and Magnitude Constraints

Paul Magron

Université de Lorraine, CNRS, Inria, LORIA
Nancy, France
paul.magron@inria.fr

Tuomas Virtanen

Audio Research Group, Tampere University
Tampere, Finland
tuomas.virtanen@tuni.fi

Abstract—Audio source separation is often achieved by estimating the magnitude spectrogram of each source, and then applying a phase recovery (or *spectrogram inversion*) algorithm to retrieve time-domain signals. Typically, spectrogram inversion is treated as an optimization problem involving one or several terms in order to promote estimates that comply with a consistency property, a mixing constraint, and/or a target magnitude objective. Nonetheless, it is still unclear which set of constraints and problem formulation is the most appropriate in practice. In this paper, we design a general framework for deriving spectrogram inversion algorithm, which is based on formulating optimization problems by combining these objectives either as soft penalties or hard constraints. We solve these by means of algorithms that perform alternating projections on the subsets corresponding to each objective/constraint. Our framework encompasses existing techniques from the literature as well as novel algorithms. We investigate the potential of these approaches for a speech enhancement task. In particular, one of our novel algorithms outperforms other approaches in a realistic setting where the magnitudes are estimated beforehand using a neural network.

Index Terms—Audio source separation, spectrogram inversion, phase recovery, alternating projections, speech enhancement.

I. INTRODUCTION

Audio source separation consists in extracting the underlying *sources* that add up to form an observed audio *mixture*. Typical approaches for source separation consist in using a deep neural network (DNN) to estimate a nonnegative mask that is applied to a time-frequency (TF) representation of the audio mixture, such as the short-time Fourier transform (STFT) [1]. Alternatively, some works operate in the time domain directly, where the STFT is replaced with learned filterbanks [2], [3]. Nonetheless, it was recently shown that TF approaches remain interesting, since they yield competitive results with lighter and more interpretable networks [4], [5].

Nonnegative TF masking results in assigning the mixture’s STFT phase to each isolated source, which induces residual interference and artifacts in the estimates. Consequently, a significant research effort has been put on phase recovery, also called *spectrogram inversion*, in DNN-based source separation, whether via deep phase models [6], or via iterative phase recovery algorithms that can be applied as a post-processing [7] or unfolded within end-to-end systems [8], [9].

Iterative spectrogram inversion algorithms are usually derived as solutions to optimization problems involving one or several terms that promote desirable properties in the TF

domain. For instance, the multiple input spectrogram inversion (MISI) algorithm [7] minimizes a measure of magnitude spectrogram mismatch under a mixing constraint (the estimates must add up to the mixture). Alternatively, the authors in [10] relax the mixing constraint into a soft penalty that is added to a *consistency* [11] term. Conversely, [12] considers a hard magnitude constraint and discards the consistency criterion. Nonetheless, it is still unclear which set of constraints and problem formulation is the most appropriate in practice.

In this paper, we design a general framework for deriving spectrogram inversion algorithms. We consider three objectives in the TF domain: *mixing*, *consistency*, and *magnitude match*, and we formulate optimization problems by combining these objectives either as soft penalties or hard constraints. We then derive an auxiliary function for each term, which allows to solve the corresponding optimization problems by means of alternating projection algorithms. While this framework encompasses existing techniques from the literature, it also allows to derive novel algorithms. We experimentally assess the potential of these algorithms for a speech enhancement task on an freely available audio corpus. In particular, one of our novel algorithms outperforms baseline MISI variants [7], [10] in a realistic setting where the magnitudes are estimated beforehand using a DNN.

The rest of this paper is structured as follows. Section II introduces the proposed framework, and algorithms are derived in Section III. Section IV presents the experimental results. Finally, Section V concludes the paper.

II. PROPOSED FRAMEWORK

A. Problem setting

Let us consider a monaural instantaneous mixture model:

$$\mathbf{X} = \sum_{j=1}^J \mathbf{S}_j, \quad (1)$$

where $\mathbf{X} \in \mathbb{C}^{F \times T}$ is the mixture STFT, and $\mathbf{S}_j \in \mathbb{C}^{F \times T}$ are the J sources matrices, whose entries are denoted $x_{f,t}$ and $s_{j,f,t}$. F and T respectively denote the number of frequency channels and time frames of the STFT. We assume that the sources’ STFT magnitudes \mathbf{V}_j have been estimated beforehand, e.g., via a DNN. Then, spectrogram inversion consists in estimating the complex-valued sources $\mathbf{S} = \{\mathbf{S}_1, \dots, \mathbf{S}_J\}$

in order to further invert these for retrieving time-domain signals. To perform this task, we search for source estimates that comply with the following properties:

- **Mixing:** the estimates should sum up to the mixture according to (1), so that there is no creation or destruction of energy overall. Such estimates are said to be *conservative*.
- **Consistency:** the estimates should be *consistent* [11], that is, the corresponding complex-valued matrices should be the STFT of time-domain signals.
- **Magnitude match:** the magnitudes of the estimates should remain close to the target magnitudes \mathbf{V}_j that have been estimated beforehand.

We propose to define a loss function corresponding to each objective in an optimization framework, and to combine these either as soft penalties or hard constraints. To solve the corresponding optimization problems, we resort to the auxiliary function method, which has shown powerful for spectrogram inversion [11]–[13]. In a nutshell, if we consider minimization of a function ϕ with parameters θ , this approach consists in constructing and minimizing an *auxiliary function* ϕ^+ with additional *auxiliary* parameters $\tilde{\theta}$ such that $\forall \theta, \phi(\theta) = \min_{\tilde{\theta}} \phi^+(\theta, \tilde{\theta})$. Then, it can easily be shown that ϕ is non-increasing under the following update scheme:

$$\tilde{\theta} \leftarrow \arg \min_{\tilde{\theta}} \phi^+(\theta, \tilde{\theta}) \quad \text{and} \quad \theta \leftarrow \arg \min_{\theta} \phi^+(\theta, \tilde{\theta}). \quad (2)$$

B. Mixing error

We consider the following mixing error:

$$h(\mathbf{S}) = \|\mathbf{X} - \sum_j \mathbf{S}_j\|^2, \quad (3)$$

where $\|\cdot\|$ is the Frobenius norm. Let us consider auxiliary parameters $\mathbf{Y} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_J\}$ such that $\sum_j \mathbf{Y}_j = \mathbf{X}$, and nonnegative weights $\Lambda_j = \{\lambda_{j,f,t}\}_{f,t}$ such that $\sum_j \lambda_{j,f,t} = 1$. We define h^+ as follows:

$$h^+(\mathbf{S}, \mathbf{Y}) = \sum_{j,f,t} \frac{|y_{j,f,t} - s_{j,f,t}|^2}{\lambda_{j,f,t}}. \quad (4)$$

Then one can show using the Jensen inequality that [12]:

$$h(\mathbf{S}) = \min_{\mathbf{Y}} h^+(\mathbf{S}, \mathbf{Y}) \quad \text{s. t.} \quad \sum_j \mathbf{Y}_j = \mathbf{X}, \quad (5)$$

which shows that h^+ is an auxiliary function for h . Besides, the auxiliary parameters' update is given by [12]:

$$\mathbf{Y}_j = \mathbf{S}_j + \Lambda_j \odot (\mathbf{X} - \sum_k \mathbf{S}_k), \quad (6)$$

where \odot denotes the element-wise matrix multiplication.

C. Inconsistency

To promote consistent estimates, we consider the inconsistency measure defined as [11]:

$$i(\mathbf{S}) = \sum_j \|\mathbf{S}_j - \mathcal{G}(\mathbf{S}_j)\|^2, \quad (7)$$

where $\mathcal{G} = \text{STFT} \circ \text{iSTFT}$. It is proven [11] that $\mathcal{G}(\mathbf{S}_j)$ is the closest consistent matrix to \mathbf{S}_j in a least-square sense, that is:

$$\|\mathbf{S}_j - \mathcal{G}(\mathbf{S}_j)\|^2 = \min_{\mathbf{Z}_j} \|\mathbf{S}_j - \mathbf{Z}_j\|^2 \quad \text{s. t.} \quad \mathbf{Z}_j \in \text{Im}(\text{STFT}), \quad (8)$$

where $\text{Im}(\text{STFT})$ is the image set of the STFT operator. Therefore, $i^+(\mathbf{S}, \mathbf{Z}) = \sum_j \|\mathbf{S}_j - \mathbf{Z}_j\|^2$ is an auxiliary function for i , and the auxiliary parameters' update is given by:

$$\mathbf{Z}_j = \mathcal{G}(\mathbf{S}_j). \quad (9)$$

D. Magnitude mismatch

Finally, we consider the following loss for characterizing the magnitude mismatch:¹

$$m(\mathbf{S}) = \sum_j \||\mathbf{S}_j| - \mathbf{V}_j\|^2. \quad (10)$$

We introduce a set of auxiliary parameters \mathbf{U}_j such that $|\mathbf{U}_j| = \mathbf{V}_j$. Then, drawing on [13], we have

$$\||\mathbf{S}_j| - \mathbf{V}_j\|^2 = \min_{\mathbf{U}_j} \|\mathbf{S}_j - \mathbf{U}_j\|^2 \quad \text{s. t.} \quad |\mathbf{U}_j| = \mathbf{V}_j, \quad (11)$$

and the minimum is reached for:

$$\mathbf{U}_j = \frac{\mathbf{S}_j}{|\mathbf{S}_j|} \odot \mathbf{V}_j. \quad (12)$$

This proves that $m^+(\mathbf{S}, \mathbf{U}) = \sum_j \|\mathbf{S}_j - \mathbf{U}_j\|^2$ is an auxiliary function for m .

III. ALGORITHMS DERIVATION

In this section, we derive spectrogram inversion algorithms. For a clarity purpose, we focus on formulating the optimization problems and providing the updates, while the complete mathematical derivations are detailed in a supporting document.²

A. Mixing and consistency as objectives

First, let us ignore the magnitude constraint, and consider the following objective function:

$$\min_{\mathbf{S}} h(\mathbf{S}) + \sigma i(\mathbf{S}), \quad (13)$$

where $\sigma \geq 0$ is a weight adjusting the relative importance of the consistency constraint (for notation purposes, $\sigma = +\infty$ corresponds to an inconsistency objective only). Using our proposed framework, the problem rewrites:

$$\begin{aligned} & \min_{\mathbf{S}, \mathbf{Y}, \mathbf{Z}} h^+(\mathbf{S}, \mathbf{Y}) + \sigma i^+(\mathbf{S}, \mathbf{Z}) \\ & \text{s. t.} \quad \sum_j \mathbf{Y}_j = \mathbf{X} \quad \text{and} \quad \mathbf{Z}_j \in \text{Im}(\text{STFT}). \end{aligned} \quad (14)$$

The updates for \mathbf{Y} and \mathbf{Z} have been derived previously (see Section II-B and II-C), thus we only need to obtain the update for \mathbf{S} . To do so, we set the partial derivative of the objective function in (14) with respect to \mathbf{S}_j at 0 and solve, which yields:

$$\mathbf{S}_j = \frac{\mathbf{Y}_j + \sigma \Lambda_j \odot \mathbf{Z}_j}{1 + \sigma \Lambda_j}, \quad (15)$$

¹Note that we recently investigated alternative magnitude discrepancy measures [14], but we focus on the Frobenius norm in this study.

²https://magronp.github.io/files/2023specinv_sup.pdf

where division is meant element-wise. Therefore, alternating updates (6), (9), and (15) yields an iterative procedure that solves (13). We call it `Mix+Incons`, and we remark that:

- This procedure is similar to the first version of the consistent Wiener filtering [11], which forces the estimates to be close to Wiener filter estimates instead of \mathbf{Y}_j . Nonetheless, both sets of estimates are conservative.
- Choosing $\sigma = 0$ or $\sigma = +\infty$ leads to $\mathbf{S}_j = \mathbf{Y}_j$ and $\mathbf{S}_j = \mathbf{Z}_j$, respectively. These non-iterative estimators are termed “mixture-consistent projection” and “STFT-consistent projection” in [15]. Thus, the general update given by (15) allows for a smooth trade-off between these.

B. Mixing and consistency with a hard magnitude constraint

We now incorporate an additional hard magnitude constraint into (13) by means of the method of Lagrange multipliers. This results in finding a critical point for:

$$h^+(\mathbf{S}, \mathbf{Y}) + \sigma i^+(\mathbf{S}, \mathbf{Z}) + \sum_{j,f,t} \delta_{j,f,t} (|s_{j,f,t}|^2 - v_{j,f,t}^2), \quad (16)$$

where $\{\delta_{j,f,t}\}_{j,f,t}$ are the Lagrange multipliers. We set the partial derivative of (16) with respect to \mathbf{S}_j at 0 and solve, which yields:

$$\mathbf{S}_j = \frac{\mathbf{Y}_j + \sigma \Lambda_j \odot \mathbf{Z}_j}{|\mathbf{Y}_j + \sigma \Lambda_j \odot \mathbf{Z}_j|} \odot \mathbf{V}_j. \quad (17)$$

We call this procedure `Mix+Incons_hardMag`. Note that:

- This procedure is equivalent to the modified MISI algorithm from [10], which however treats the weights Λ as unknown parameters and updates them at each iteration.
- If $\sigma = +\infty$, the procedure boils down to applying the well-known Griffin-Lim update [16] to each source independently without mixing constraint.
- If $\sigma = 0$, the procedure reduces to our previous “PU-Iter” [12], which discards the consistency constraint.

C. Consistency objective with a hard mixing constraint

Now, let us consider an inconsistency-only objective function, where mixing is treated as a hard constraint. Note that in this setup, we do not consider an additional hard magnitude constraint since this yields an ill-posed problem.³ As above, we treat this problem with the method of Lagrange multipliers, which eventually yields:

$$\mathbf{S}_j = \mathbf{Z}_j + \frac{1}{J} (\mathbf{X} - \sum_k \mathbf{Z}_k), \quad (18)$$

where the update for \mathbf{Z} is given by (9). We call this method `Incons_hardMix`, and we remark that:

- This approach is non-iterative, since the set of consistent matrices is a vector space and \mathbf{Z}_j is consistent by construction.
- It is equivalent to the successive application of the STFT- and mixture-consistent projections used in [15].

³Indeed, one can verify on a simple example ($J = 2$, $v_1 = v_2 = 1$, and $x = 4$) that there is no solution that satisfies both constraints in general.

- The update (18) is similar to (6) with fixed weights $\Lambda_j = 1/J$, which is expected when using mixing as a hard constraint [7].

D. Magnitude objective with a hard mixing constraint

Finally, we consider the magnitude mismatch as the main objective under a hard mixing constraint. Since incorporating an additional hard consistency constraint would eventually yield MISI [13], we focus here on a soft consistency penalty:

$$\min_{\mathbf{S}} m(\mathbf{S}) + \sigma i(\mathbf{S}) \quad \text{s. t.} \quad \sum_j \mathbf{S}_j = \mathbf{X}. \quad (19)$$

Still using the method of Lagrange multipliers, we obtain:

$$\mathbf{W}_j = \frac{1}{1 + \sigma} (\mathbf{U}_j + \sigma \mathbf{Z}_j), \quad (20)$$

$$\mathbf{S}_j = \mathbf{W}_j + \frac{1}{J} \left(\mathbf{X} - \sum_k \mathbf{W}_k \right), \quad (21)$$

where \mathbf{Z}_j and \mathbf{U}_j are given by (9) and (12). We call this procedure `Mag+Incons_hardMix`.

Remark: Let us note that if we discard the consistency constraint ($\sigma = 0$) and initialize the estimates using an amplitude mask (see Section IV-A), then the estimator becomes:

$$\mathbf{S}_j = \left(\mathbf{V}_j + \frac{1}{J} (|\mathbf{X}| - \sum_k \mathbf{V}_k) \right) \frac{\mathbf{X}}{|\mathbf{X}|}, \quad (22)$$

which is non-iterative and assigns the mixture’s phase to each source, therefore it is not interesting for a phase recovery purpose.

E. Summary of the algorithms

We summarize in Table I the updates performed by these various algorithms using the following magnitude, consistency, and mixing projectors:

$$\mathcal{P}_{\text{mag}}(\mathbf{S}) = \left\{ \frac{\mathbf{S}_j}{|\mathbf{S}_j|} \odot \mathbf{V}_j \right\}_j \quad (23)$$

$$\mathcal{P}_{\text{cons}}(\mathbf{S}) = \{\mathcal{G}(\mathbf{S}_j)\}_j \quad (24)$$

$$\mathcal{P}_{\text{mix}}(\mathbf{S}) = \left\{ \mathbf{S}_j + \Lambda_j \odot (\mathbf{X} - \sum_k \mathbf{S}_k) \right\}_j. \quad (25)$$

IV. EXPERIMENTS

In this section, we assess the potential of our algorithms for a speech enhancement task, a particular case of source separation with $J = 2$ sources (speech and noise). Note that this framework remains applicable to other source separation scenarios such as speech [1] or music [17] separation. Our code is available online.⁴

⁴<https://github.com/magronp/spectrogram-inversion>

TABLE I: Alternating projection algorithms for spectrogram inversion. Λ_j denotes mixing weights that can be hand-tuned.

Algorithm	Reference	Consistency weight	Mixing weights	Iterative	Update formula
MISI	[7]	no	$1/J$	yes	$\mathcal{P}_{\text{mix}}(\mathcal{P}_{\text{mag}}(\mathcal{P}_{\text{cons}}(\mathbf{S})))$
Mix+Incons		σ	$\{\Lambda_j\}_j$	yes	$\frac{1}{1 + \sigma\Lambda} \odot (\mathcal{P}_{\text{mix}}(\mathbf{S}) + \sigma\Lambda \odot \mathcal{P}_{\text{cons}}(\mathbf{S}))$
Mixture-consistent projection	[15]	$\sigma = 0$		no	$\mathcal{P}_{\text{mix}}(\mathbf{S})$
STFT-consistent projection	[15]	$\sigma = +\infty$		no	$\mathcal{P}_{\text{cons}}(\mathbf{S})$
Mix+Incons_hardMag		σ	$\{\Lambda_j\}_j$	yes	$\mathcal{P}_{\text{mag}}(\mathcal{P}_{\text{mix}}(\mathbf{S}) + \sigma\Lambda \odot \mathcal{P}_{\text{cons}}(\mathbf{S}))$
Modified MISI	[10]	σ	optimized	yes	$\mathcal{P}_{\text{mag}}(\mathcal{P}_{\text{mix}}(\mathbf{S}) + \sigma\Lambda \odot \mathcal{P}_{\text{cons}}(\mathbf{S}))$
PU-Iter	[12]	$\sigma = 0$	$\{\Lambda_j\}_j$	yes	$\mathcal{P}_{\text{mag}}(\mathcal{P}_{\text{mix}}(\mathbf{S}))$
Incons_hardMix	[15]	no	$1/J$ or $\{\Lambda_j\}_j$ or learned	no	$\mathcal{P}_{\text{mix}}(\mathcal{P}_{\text{cons}}(\mathbf{S}))$
Mag+Incons_hardMix		σ	$1/J$	yes	$\mathcal{P}_{\text{mix}}\left(\frac{1}{1 + \sigma}(\mathcal{P}_{\text{mag}}(\mathbf{S}) + \sigma\mathcal{P}_{\text{cons}}(\mathbf{S}))\right)$

A. Protocol

Data: As acoustic material, we build mixtures of clean speech and noise. We randomly select 100 utterances from the VoiceBank set [18] to create the clean speech, and we select three real-world environments noise signals (living room, bus, and public square) from the DEMAND dataset [19]. For each clean speech signal, we randomly select a noise excerpt cropped at the same length than that of the speech signal. We then mix the two signals at various input signal-to-noise ratios (iSNRs) (10, 0, and -10 dB). All audio excerpts are single-channel and sampled at 16 kHz. The STFT is computed with a 1024 samples-long (64 ms) Hann window and 75% overlap. The dataset is split into two subsets of 50 mixtures: a *validation* set, on which the consistency weight is tuned; and a *test* set, on which all algorithms are evaluated.

Spectrogram estimation: We estimate the magnitude spectrograms \mathbf{V}_j with Open-Unmix [17], an open implementation of a three-layer BLSTM neural network, originally tailored for music source separation, and later adapted to speech enhancement. We use the pre-trained model available at [20] and described in [18]. In practice, magnitudes are estimated more accurately as the iSNR increases.

Methods: All algorithms are initialized with an amplitude mask (AM), i.e., the estimated magnitudes are combined with the mixture’s phase. The number of iterations is tuned on the validation set (see next section), with a maximum of 20. For the mixing projector, any nonnegative weights Λ_j that verify the sum-to-one constraint can be used. In practice, we consider magnitude ratios $\Lambda_j = \mathbf{V}_j / \sum_k \mathbf{V}_k$, since these are common in such algorithms [12], [15] and yield the same performance as the optimized weights described in III-B.

Metric: We evaluate the separation quality via the signal-to-distortion ratio (SDR) between the true clean speech \mathbf{s}_1^* and its estimate \mathbf{s}_1 (averaged over signals, higher is better):

$$\text{SDR}(\mathbf{s}_1^*, \mathbf{s}_1) = 20 \log_{10} \frac{\|\mathbf{s}_1^*\|}{\|\mathbf{s}_1^* - \mathbf{s}_1\|}. \quad (26)$$

B. Results

First, let us investigate the influence of the consistency weight σ onto performance. From the results displayed in

TABLE II: Test results (SDR in dB).

	iSNR= 10	iSNR= 0	iSNR= -10
AM	18.7	13.5	7.7
MISI	19.6	14.1	7.7
Mix+Incons	19.3	13.7	8.1
Mix+Incons_hardMag	18.7	13.8	7.9
Incons_hardMix	19.6	13.9	7.5
Mag+Incons_hardMix	19.6	14.1	7.7

Fig. 1a-1c, we remark that Mag+Incons_hardMix exhibits a stable performance with respect to σ . Conversely, for the other algorithms, the SDR peaks at specific values of σ . In particular, a properly tuned Mix+Incons outperforms its special cases corresponding to $\sigma = 0$ and $+\infty$, which were used in [15]. A similar behavior is observed for Mix+Incons_hardMag, which outperforms its special case $\sigma = 0$ [12]. This demonstrates the interest of our framework, which allows for deriving general algorithms that outperform their special cases found in the literature.

Besides, Fig. 1d reveals that the algorithms exhibit a different behavior over iterations. Indeed, while for most methods, the SDR steadily increases over iterations, MISI reaches its peak performance after very few iterations, and then its performance drops severely. This was notably observed in [10]; nonetheless, the modified MISI algorithm introduced in this paper (which is similar to Mix+Incons_hardMag) was compared to MISI after 200 iterations, which is somewhat unfair to MISI. Here, we select the optimal weight and number of iterations for each algorithm in order to run them on the test set in a more fair setup.

From the test results presented in Table II, we remark that MISI achieves the best performance at high or moderate iSNR, i.e., when the spectrograms are rather accurately estimated. We also observe that Mag+Incons_hardMix can be an interesting alternative to MISI: indeed, both algorithms perform similarly in terms of SDR, but the latter is easier to tune, as is evident from its steady behavior over iterations and stability with respect to σ (see Fig. 1). The Incons_hardMix algorithm, which was used for end-to-end source separation in [15], reveals interesting at high iSNR since it is non-iterative

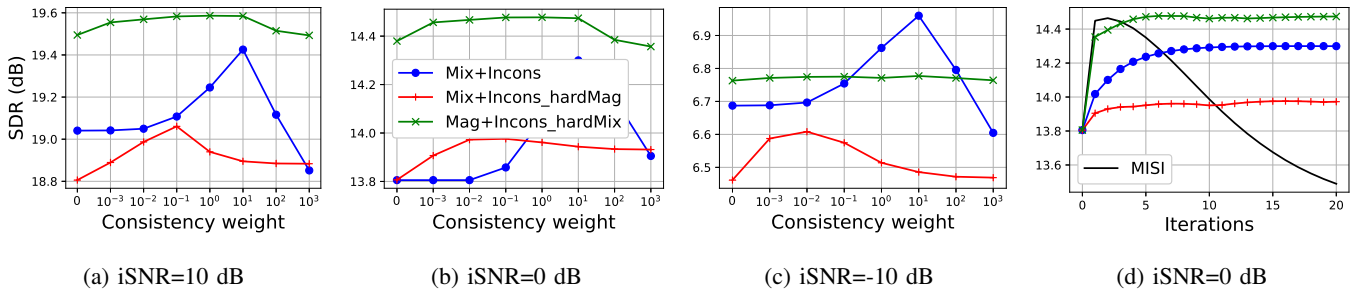


Fig. 1: Validation SDR as a function of σ for the optimal number of iterations at various iSNRs (1a-1c), and validation SDR over iterations for the optimal consistency weight at iSNR=0 dB (1d); similar trends are observed at other iSNRs.

and yields similar results to MISI. However, it performs the worst at low iSNR, where a baseline AM is preferable. On the other hand, while `Mix+Incons_hardMag` improves over MISI at low iSNR, it becomes less interesting when spectrograms are more accurately estimated (iSNR = 0 or 10 dB), which differs from the results of [10]. This difference might be explained by the usage of a different magnitude estimation technique which is of paramount importance in phase recovery [12] (spectral subtraction in [10] vs. a DNN here); and by the afore-mentioned impact of an optimized number of iterations. Finally, we note that the proposed `Mix+Incons` allows to mitigate this SDR drop at high iSNR, while it further improves the performance at low iSNR by 0.2 dB over the previously best performing approach. It should be noted that this technique does not exploit the magnitude projector, and only relies on the estimated magnitudes V_j via its initialization. Therefore, the algorithm allows for deviation from these magnitude values, which might explain its good performance at low iSNR, since magnitude are estimated with a lower accuracy in this case.

V. CONCLUSION

In this paper, we introduced a general framework for deriving alternating projection algorithms for spectrogram inversion using consistency, mixing, and magnitude constraints. This framework encompasses existing techniques from the literature, but also yields novel algorithms, among which some appear as promising alternatives to baseline spectrogram inversion approaches. Future work will be devoted to adapt these algorithms to operate online [13] in order to combine them with model-based phase priors [6], [12]. We will also unfold them into deep networks for end-to-end separation, where supervised learning can be leveraged to optimize the consistency and mixing weights.

REFERENCES

- [1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
- [2] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
- [3] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Filterbank design for end-to-end speech separation," in *Proc. IEEE ICASSP*, May 2020.
- [4] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *Proc. IEEE ICASSP*, May 2020.
- [5] T. Cord-Landwehr, C. Boeddeker, T. Von Neumann, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach, "Monaural source separation: From anechoic to reverberant environments," in *Proc. IWAENC*, Sept. 2022.
- [6] Y. Masuyama, K. Yatabe, K. Nagatomo, and Y. Oikawa, "Online phase reconstruction via DNN-based phase differences estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 163–176, 2023.
- [7] 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.
- [8] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, "End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction," in *Proc. Interspeech*, Sept. 2018.
- [9] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Deep Griffin-Lim iteration: Trainable iterative phase reconstruction using neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 37–50, Jan. 2021.
- [10] D. Wang, H. Kameoka, and K. Shinoda, "A modified algorithm for multiple input spectrogram inversion," in *Proc. Interspeech*, Sept. 2019.
- [11] 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*, Sept. 2010.
- [12] P. Magron, R. Badeau, and B. David, "Model-based STFT phase recovery for audio source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 6, pp. 1095–1105, Jun. 2018.
- [13] P. Magron and T. Virtanen, "Online spectrogram inversion for low-latency audio source separation," *IEEE Signal Processing Letters*, vol. 27, pp. 306–310, 2020.
- [14] P. Magron, P.-H. Vial, T. Oberlin, and C. Févotte, "Phase recovery with Bregman divergences for audio source separation," in *Proc. IEEE ICASSP*, Jun. 2021.
- [15] S. Wisdom, J. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. IEEE ICASSP*, May 2019.
- [16] 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, Apr. 1984.
- [17] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - a reference implementation for music source separation," *Journal of Open Source Software*, 2019.
- [18] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Proc. Interspeech*, Sept. 2016.
- [19] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," <https://doi.org/10.5281/zenodo.1227121>, Jun. 2013.
- [20] S. Uhlich and Y. Mitsufuji, "Open-unmix for speech enhancement (UMX SE)," <https://doi.org/10.5281/zenodo.3786908>, May 2020.