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

Supplementary Material

Paul Magron, Tuomas Virtanen*

Abstract

This document provides additional mathematical derivations to the main paper [1], which addresses the problem of spectrogram inversion for audio source separation. We briefly recall the problem, the considered time-frequency constraints (consistency, mixing, and magnitude match), and the optimization framework (losses and auxiliary functions). We then derive alternating projection algorithms for solving the corresponding problems.

1 Preliminaries

1.1 Mathematical notations

- $\mathbf{A}$ (capital, bold font): matrix, whose (i,j)-th entry is denoted $a_{i,j}$.
- z (regular): scalar.
- $|.|$, $\angle(.)$: magnitude and complex angle, respectively.
- $\bar{.}$: complex conjugate.
- $\Re(.)$: real part function.
- $\overset{c}{=}$: equality up to an additive constant.
- Im(.): image set of an operator.
- $||.||$: Frobenius norm.
- $\odot$, fraction bar: element-wise matrix multiplication and division, respectively.

1.2 Problem setting

We consider the source separation problem in the STFT domain, and use the following monaural instantaneous mixture model:

$$\mathbf{X} = \sum_{j=1}^{J} \mathbf{S}_j, \tag{1}$$

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. 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, thus our goal is to estimate complex-valued sources $\mathbf{S}_j$, in order to further invert these STFTs for retrieving time-domain signals.

*P. Magron is with Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France. T. Virtanen is with Audio Research Group, Tampere University, Tampere, Finland.

1.3 Constraints

To perform this task, we search for source estimates that satisfy 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*, 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.

1.4 Auxiliary function method

We define a loss function corresponding to each objective in an optimization framework, and we combine these either as soft penalties or hard constraints.

To solve the corresponding optimization problems, we resort to the auxiliary function method. 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 $\forall \theta, \quad \phi(\theta) = \min_{\tilde{\theta}} \phi^+(\theta, \tilde{\theta})$. Then, it can easily be shown that ϕ is non-increasing under the following update scheme:

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

2 Objectives and auxiliary functions

2.1 Mixing error

As defined in the main document, the mixing error is:

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

and we consider the following auxiliary function:

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

with auxiliary parameters $\mathbf{Y}_j$ such that $\sum_j \mathbf{Y}_j = \mathbf{X}$, and nonnegative weights $\mathbf{\Lambda}_j$, such that $\sum_j \lambda_{j,f,t} = 1$ for all f, t. To obtain the update for the auxiliary parameters $\mathbf{Y}$, we introduce the constraint using the method of Lagrange multipliers which leads to finding a critical point of the following functional:

$$h^+(\mathbf{S}, \mathbf{Y}) + 2\Re\left(\sum_{f,t} \gamma_{f,t}(x_{f,t} - \sum_j y_{j,f,t})\right). \tag{5}$$

To that end, we set its partial derivative with respect to (w.r.t.) $\mathbf{Y}_j$ at 0, which leads to:

$$\frac{1}{\lambda_{j,f,t}}(y_{j,f,t} - s_{j,f,t}) - \gamma_{f,t} = 0. \tag{6}$$

Summing (6) over j and using $\sum_j \lambda_{j,f,t} = 1$ and $\sum_j \mathbf{Y}_j = \mathbf{X}$ allows to determine the multipliers $\gamma_{f,t}$

$$\gamma_{f,t} = x_{j,f,t} - \sum_j s_{j,f,t}, \tag{7}$$

which finally leads to the update rule:

$$\mathbf{Y}_j = \mathbf{S}_j + \mathbf{\Lambda}_j \odot (\mathbf{X} - \sum_k \mathbf{S}_k). \tag{8}$$

2.2 Inconsistency

Inconsistency is defined as follows:

$$i(\mathbf{S}) = \sum_j ||\mathbf{S}_j - \mathcal{G}(\mathbf{S}_j)||^2, \tag{9}$$

where $\mathcal{G} = \text{STFT} \circ \text{iSTFT}$. As pointed out in the main paper, we have:

$$||\mathbf{S}_j - \mathcal{G}(\mathbf{S}_j)||^2 = \min_{\mathbf{Z}_j \in \text{Im(STFT)}} ||\mathbf{S}_j - \mathbf{Z}_j||^2, \tag{10}$$

then $i^+(\mathbf{S}, \mathbf{Z}) = \sum_j ||\mathbf{S}_j - \mathbf{Z}_j||^2$ is an auxiliary function for i, and it is minimized w.r.t. $\mathbf{Z}$ through the update:

$$\mathbf{Z}_j = \mathcal{G}(\mathbf{S}_j). \tag{11}$$

2.3 Magnitude mismatch

Finally, the magnitude loss is:

$$m(\mathbf{S}) = \sum_{j,f,t} ||s_{j,f,t}| - v_{j,f,t}|^2. \tag{12}$$

We introduce a set of auxiliary parameters $\mathbf{U}_j$ such that $|\mathbf{U}_j| = \mathbf{V}_j$, and rewrite the cost function using:

$$||s_{j,f,t}| - v_{j,f,t}|^2 = ||s_{j,f,t}| - |u_{j,f,t}||^2 \leq |s_{j,f,t} - u_{j,f,t}|^2, \tag{13}$$

where we use the fact that $\forall (z, z') \in \mathbb{C}^2$, $||z| - |z'|| \leq |z - z'|$ and where the equality holds if $\angle z = \angle z'$ (this is easy to demonstrate using the triangle inequality). This leads to:

$$m(\mathbf{S}) \leq \sum_{j,f,t} |s_{j,f,t} - u_{j,f,t}|^2. \tag{14}$$

In other words, $m(\mathbf{S}) \leq m^+(\mathbf{S}, \mathbf{U}_j)$ with:

$$m^+(\mathbf{S}, \mathbf{Y}) = \sum_j ||\mathbf{S}_j - \mathbf{U}_j||^2, \tag{15}$$

To minimize m^+ w.r.t. $\mathbf{U}$ under the constraint $|\mathbf{U}_j| = \mathbf{V}_j$, we again resort to the method of Lagrange multipliers. We therefore aim at finding a critical point for:

$$m^+(\mathbf{S}, \mathbf{U}) + \sum_{j,f,t} \delta_{j,f,t}(|u_{j,f,t}|^2 - v_{j,f,t}^2), \tag{16}$$

where $\delta_{j,f,t}$ are Lagrange multipliers. We set the partial derivative of (16) w.r.t $\mathbf{U}_j$ at 0 which yields:

$$y_{j,f,t} - s_{j,f,t} + \delta_{j,f,t} u_{j,f,t} = 0, \tag{17}$$

which rewrites:

$$(1 + \delta_{j,f,t}) u_{j,f,t} = s_{j,f,t}. \tag{18}$$

Using the constraint $|u_{j,f,t}| = v_{j,f,t}$, we have:

$$|1 + \delta_{j,f,t}| = \frac{|s_{j,f,t}|}{v_{j,f,t}}. \tag{19}$$

which finally leads to the update for $\mathbf{U}_j$:

$$\mathbf{U}_j = \pm \frac{\mathbf{S}_j}{|\mathbf{S}_j|} \odot \mathbf{V}_j. \tag{20}$$

We retain the update that do not modify the phase of $\mathbf{S}_j$, since it ensures that $m(\mathbf{S}) = m^+(\mathbf{S}, \mathbf{U}_j)$, which shows that m^+ is an auxiliary function for m.

3 Algorithms derivation

3.1 `Mix+Incons`

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

$$\min_{\mathbf{S}} h(\mathbf{S}) + \sigma i(\mathbf{S}), \tag{21}$$

where $\sigma \geq 0$ is a weight adjusting the relative importance of the consistency constraint. Using the previously derived auxiliary function for each term, the problem rewrites:

$$\min_{\mathbf{S},\mathbf{Y},\mathbf{Z}} h^+(\mathbf{S},\mathbf{Y}) + \sigma i^+(\mathbf{S},\mathbf{Z}) \quad \text{s. t.} \quad \sum_j \mathbf{Y}_j = \mathbf{X} \quad \text{and} \quad \mathbf{Z}_j \in \mathrm{Im}(\mathrm{STFT}). \tag{22}$$

The updates for $\mathbf{Y}$ and $\mathbf{Z}$ have already been obtained (*cf.* (8) and (11)), so we only need to obtain the update for $\mathbf{S}$. To do so, we set the partial derivative of the objective function in (22) w.r.t. $\mathbf{S}_j$ at 0:

$$\frac{\mathbf{S}_j - \mathbf{Y}_j}{\mathbf{\Lambda}_j} + \sigma(\mathbf{S}_j - \mathbf{Z}_j) = 0, \tag{23}$$

and solving yields:

$$\mathbf{S}_j = \frac{\mathbf{Y}_j + \sigma \mathbf{\Lambda}_j \odot \mathbf{Z}_j}{1 + \sigma \mathbf{\Lambda}_j}. \tag{24}$$

3.2 `Mix+Incons_hardMag`

Now, we incorporate a hard magnitude constraint ($\forall j$, $|\mathbf{S}_j| = \mathbf{V}_j$) in (22) by means of the method of Lagrange multipliers. We therefore aim at finding a critical point for:

$$\mathcal{H}(\mathbf{S},\mathbf{Y},\mathbf{Z},\boldsymbol{\delta}) = h^+(\mathbf{S},\mathbf{Y}) + \sigma i^+(\mathbf{S},\mathbf{Z}) + \sum_{j,f,t} \delta_{j,f,t}(|s_{j,f,t}|^2 - v_{j,f,t}^2), \tag{25}$$

where we discarded the constraints that only refer to $\mathbf{Y}$ and $\mathbf{Z}$, since their updates are already known. We set the partial derivative of $\mathcal{H}$ w.r.t $\mathbf{S}_j$ at 0:

$$\frac{1}{\lambda_{j,f,t}}(s_{j,f,t} - y_{j,f,t}) + \sigma(s_{j,f,t} - z_{j,f,t}) + \delta_{j,f,t} s_{j,f,t} = 0, \tag{26}$$

which rewrites:

$$(1 + (\delta_{j,f,t} + \sigma)\lambda_{j,f,t}) s_{j,f,t} = y_{j,f,t} + \sigma \lambda_{j,f,t} z_{j,f,t}. \tag{27}$$

Using the constraint $|s_{j,f,t}| = v_{j,f,t}$, we have:

$$1 + (\delta_{j,f,t} + \sigma)\lambda_{j,f,t} = \frac{|y_{j,f,t} + \sigma \lambda_{j,f,t} z_{j,f,t}|}{v_{j,f,t}}, \tag{28}$$

which finally leads to the update for $\mathbf{S}_j$:

$$\mathbf{S}_j = \frac{\mathbf{Y}_j + \sigma \mathbf{\Lambda}_j \odot \mathbf{Z}_j}{|\mathbf{Y}_j + \sigma \mathbf{\Lambda}_j \odot \mathbf{Z}_j|} \odot \mathbf{V}_j. \tag{29}$$

3.3 `Incons_hardMix`

Now, let us consider that the mixing constraint is no longer a term in the objective function, which therefore reduces to the inconsistency cost only. In this setting, one can use an hard magnitude constraint, which becomes equivalent to the update (29) obtained above with $\sigma = +\infty$. Alternatively, one can consider a hard mixing constraint, which we do hereafter (as pointed out in the main paper, it is not possible to use both since they might be incompatible in practice). The problem becomes that of minimizing inconsistency (9) under the constraint $\sum_j \mathbf{S}_j = \mathbf{X}$. We seek to find a critical point for:

$$i^+(\mathbf{S},\mathbf{Z}) + 2\Re\left(\sum_{f,t} \gamma_{f,t}(x_{f,t} - \sum_j s_{j,f,t})\right). \tag{30}$$

As the update for $\mathbf{Z}$ is already known (*cf.* (11)), we set the partial derivative of (30) w.r.t. $\mathbf{S}_j$ at 0:

$$s_{j,f,t} - z_{j,f,t} - \gamma_{f,t} = 0. \tag{31}$$

Summing over j and using the hard mixing constraint, we have:

$$\gamma_{f,t} = \frac{1}{J}(x_{f,t} - \sum_j z_{j,f,t}), \tag{32}$$

which yields the update for $\mathbf{S}_j$:

$$\mathbf{S}_j = \mathbf{Z}_j + \frac{1}{J}(\mathbf{X} - \sum_k \mathbf{Z}_k). \tag{33}$$

which is actually non-iterative: indeed, since the $\mathbf{Z}_j$ are by construction consistent matrices, and since the set of consistent matrices is a vector space, then $\mathbf{S}_j$ is also consistent. Therefore, given any initial $\tilde{\mathbf{S}}_j$, the following estimate:

$$\widehat{\mathbf{S}}_j = \mathcal{G}(\tilde{\mathbf{S}}_j) + \frac{1}{J}(\mathbf{X} - \sum_k \mathcal{G}(\tilde{\mathbf{S}}_k)), \tag{34}$$

is a solution to the problem, that is, it is a zero of the inconsistency function i and it complies with the mixing constraint.

3.4 `Mag+Incons_hardMix`

We consider the magnitude mismatch as the main objective function, with a soft consistency penalty and under a hard mixing constraint.

$$\min_{\mathbf{S}} m(\mathbf{S}) + \sigma i(\mathbf{S}) \quad \text{s. t.} \quad \sum_j \mathbf{S}_j = \mathbf{X}. \tag{35}$$

3.4.1 Auxiliary function method

Using the same methodology as above results in finding a critical point for:

$$m^+(\mathbf{S}, \mathbf{U}) + \sigma i^+(\mathbf{S}, \mathbf{Z}) + 2\Re\left(\sum_{f,t} \gamma_{f,t}(x_{f,t} - \sum_j s_{j,f,t})\right). \tag{36}$$

where $\gamma_{f,t}$ are the Lagrange multipliers (again, we discard the constraints on $\mathbf{Z}$ and $\mathbf{U}$ since they are already known). Setting its partial derivative w.r.t. $\mathbf{S}_j$ at 0 yields:

$$s_{j,f,t} - u_{j,f,t} + \sigma(s_{j,f,t} - z_{j,f,t}) - \gamma_{f,t} = 0, \tag{37}$$

which rewrites:

$$(1+\sigma)s_{j,f,t} - (u_{j,f,t} + \sigma z_{j,f,t}) - \gamma_{f,t} = 0. \tag{38}$$

Summing over j and using the hard mixing constraint, we have:

$$(1+\sigma)x_{f,t} - \sum_j (u_{j,f,t} + \sigma z_{j,f,t}) - J\gamma_{f,t} = 0, \tag{39}$$

which yields:

$$\gamma_{f,t} = \frac{1}{J}\left((1+\sigma)x_{f,t} - \sum_j (u_{j,f,t} + \sigma z_{j,f,t})\right). \tag{40}$$

Incorporating this expression into (38) leads to:

$$(1+\sigma)s_{j,f,t} = (u_{j,f,t} + \sigma z_{j,f,t}) + \frac{1}{J}(1+\sigma)x_{f,t} - \sum_k (u_{k,f,t} + \sigma z_{k,f,t}). \tag{41}$$

We introduce the following notation:

$$\mathbf{W}_j = \frac{1}{1+\sigma}\left(\mathbf{U}_j + \sigma \mathbf{Z}_j\right), \tag{42}$$

which allows to write the update (41) in compact form:

$$\mathbf{S}_j = \mathbf{W}_j + \frac{1}{J}\left(\mathbf{X} - \sum_k \mathbf{W}_k\right). \tag{43}$$

Remark: Let us note that if we discard the consistency constraint ($\sigma = 0$) and initialize the estimates using an amplitude mask, i.e., $\mathbf{S}_j = \mathbf{V}_j \odot \frac{\mathbf{X}}{|\mathbf{X}|}$, then the estimator becomes:

$$\mathbf{S}_j = \left(\mathbf{V}_j + \frac{1}{J}(|\mathbf{X}| - \sum_l \mathbf{V}_l)\right)\frac{\mathbf{X}}{|\mathbf{X}|}, \tag{44}$$

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

3.4.2 Direct derivation when there is no inconsistency

Considering no consistency penalty, we can solve problem (35) directly without resorting to the auxiliary function method, that is, we aim at finding a critical point of:

$$\mathcal{M}(\mathbf{S}, \gamma, \gamma') = m(\mathbf{S}) + 2\Re\left(\sum_{f,t} \gamma_{f,t}(x_{f,t} - \sum_j s_{j,f,t})\right). \tag{45}$$

To that end, we remark that:

$$m(\mathbf{S}) \stackrel{c}{=} \sum_{j,f,t} s_{j,f,t}\bar{s}_{j,f,t} - 2v_{j,f,t}\sqrt{s_{j,f,t}\bar{s}_{j,f,t}}, \tag{46}$$

Setting the partial derivative of $\mathcal{M}$ w.r.t. $\mathbf{S}_j$ at 0 leads to (we remove the indexes f and t for clarity):

$$s_j - v_j\frac{\sqrt{s_j}}{\sqrt{\bar{s}_j}} + \gamma = 0. \tag{47}$$

Since for all $z \in \mathbb{C}^*$, $\frac{\sqrt{z}}{\sqrt{\bar{z}}} = \frac{z}{|z|}$, then. (47) rewrites:

$$s_j - v_j\frac{s_j}{|s_j|} + \gamma = 0. \tag{48}$$

Let us now write s_j in polar form: $s_j = a_j e^{\mathrm{i}\phi_j}$. We have:

$$(a_j - v_j)e^{\mathrm{i}\phi_j} + \gamma = 0. \tag{49}$$

which implies that $a_j - v_j$ and ϕ_j are constant for all j. Consequently, since $\sum_j a_j e^{\mathrm{i}\phi_j} = x$, then $\phi_j = \angle x$. Besides, since $a_j = v_j + c$, then $\sum_k v_k + Jc = |x|$, so $c = \frac{1}{J}(|x| - \sum_k v_k)$. This results in

$$s_j = \left(v_j + \frac{1}{J}(|x| - \sum_l v_l)\right)e^{\mathrm{i}\angle x}, \tag{50}$$

or equivalently:

$$s_j = \left(v_j + \frac{1}{J}(|x| - \sum_l v_l)\right)\frac{x}{|x|}. \tag{51}$$

We observe that these estimates are equivalent to those given by (44).
Remark: if we use the Kullback-Leibler divergence instead of the Euclidean distance for defining m, we obtain similar non-iterative estimates given by the amplitude ratio:

$$s_j = \frac{v_j}{\sum_k v_k}x. \tag{52}$$

Table 1: Alternating projection algorithms for spectrogram inversion.

Algorithm	Consistency	Mixing	Iterative	Update formula
MISI	no	$1/J$	yes	$\mathcal{P}_{\text{mix}}(\mathcal{P}_{\text{mag}}(\mathcal{P}_{\text{cons}}(\mathbf{S})))$
`Mix+Incons`	σ	$\{\mathbf{\Lambda}_j\}_j$	yes	$\frac{1}{1+\sigma\mathbf{\Lambda}} \odot (\mathcal{P}_{\text{mix}}(\mathbf{S}) + \sigma\mathbf{\Lambda} \odot \mathcal{P}_{\text{cons}}(\mathbf{S}))$
`Mix+Incons_hardMag`	σ	$\{\mathbf{\Lambda}_j\}_j$	yes	$\mathcal{P}_{\text{mag}}(\mathcal{P}_{\text{mix}}(\mathbf{S}) + \sigma\mathbf{\Lambda} \odot \mathcal{P}_{\text{cons}}(\mathbf{S}))$
`Incons_hardMix`	no	$1/J$	no	$\mathcal{P}_{\text{mix}}(\mathcal{P}_{\text{cons}}(\mathbf{S}))$
`Mag+Incons_hardMix`	σ	$1/J$	yes	$\mathcal{P}_{\text{mix}}\left(\frac{1}{1+\sigma}(\mathcal{P}_{\text{mag}}(\mathbf{S}) + \sigma\mathcal{P}_{\text{cons}}(\mathbf{S}))\right)$

4 Summary of the algorithms

We summarize in Table 1 the updates performed by these various MM algorithms using the following magnitude, consistency, and mixing projectors:

$$\mathcal{P}_{\text{mag}}(\mathbf{S}) = \left\{ \frac{\mathbf{S}_j}{|\mathbf{S}_j|} \odot \mathbf{V}_j \right\}_j \tag{53}$$

$$\mathcal{P}_{\text{cons}}(\mathbf{S}) = \{\mathcal{G}(\mathbf{S}_j)\}_j \tag{54}$$

$$\mathcal{P}_{\text{mix}}(\mathbf{S}) = \left\{ \mathbf{S}_j + \mathbf{\Lambda}_j \odot (\mathbf{X} - \sum_k \mathbf{S}_k) \right\}_j . \tag{55}$$

References

[1] P. Magron and T. Virtanen, "Spectrogram inversion for audio source separation via alternating projection algorithms," submitted in a conference, 2023.